Original Research Article

SMMR STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2024, Vol. 0(0) 1–19

DOI: 10.1177/09622802241238998
journals.sagepub.com/home/smm
Sage

Interval estimation in three-class receiver operating characteristic analysis: A fairly general approach based on the empirical likelihood

Duc-Khanh To[1,2], Gianfranco Adimari[3] and Monica Chiogna[4]

Abstract

The empirical likelihood is a powerful nonparametric tool, that emulates its parametric counterpart—the parametric likelihood—preserving many of its large-sample properties. This article tackles the problem of assessing the discriminatory power of three-class diagnostic tests from an empirical likelihood perspective. In particular, we concentrate on interval estimation in a three-class receiver operating characteristic analysis, where a variety of inferential tasks could be of interest. We present novel theoretical results and tailored techniques studied to efficiently solve some of such tasks. Extensive simulation experiments are provided in a supporting role, with our novel proposals compared to existing competitors, when possible. It emerges that our new proposals are extremely flexible, being able to compete with contestants and appearing suited to accommodating several distributions, such, for example, mixtures, for target populations. We illustrate the application of the novel proposals with a real data example. The article ends with a discussion and a presentation of some directions for future research.

Keywords

Bootstrap, diagnostic test, nonparametric inference, receiver operating characteristic surface, volume under the receiver operating characteristic surface

1 Introduction

The construction of confidence intervals or regions for unknown parameters of interest is a classic problem of statistical inference, and, by the time, the approach based on the empirical likelihood (EL) (see Owen,[1] as general reference) has shown its effectiveness and flexibility in addressing such kind of problem.

The EL is a nonparametric tool that allows obtaining pseudo-likelihoods in several contexts and, in particular, for parameters that are determined by estimating equations. By an emulation of its parametric counterpart, the EL function is obtained by maximization of a nonparametric likelihood supported on the data, subject to some constraints. In most cases, the maximization problem is solved by using Lagrange multipliers. This leads to an explicit expression for (minus twice) the empirical log-likelihood ratio, for which a Wilks-type theorem is generally proved. Then, the EL can be used, in a standard

[1]Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]Unit of Biostatistics, Epidemiology and Public Health—Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Italy
[4]Department of Statistical Sciences "Paolo Fortunati," University of Bologna, Italy

Corresponding author:
Duc-Khanh To, Faculty of Mathematics and Computer Science, University of Science—Vietnam National University, Nguyen Van Cu street, 227; 70000 Ho Chi Minh city, Vietnam.
Email: tdkhanh@hcmus.edu.vn

way, to obtain non-parametric confidence intervals or regions, which, by their nature, are range-respecting and have the shape determined only by data, free from any artificial constraint, such as that of symmetry. Moreover, EL regions are generally more accurate than traditional ones, based on the asymptotic normality of some estimator for the parameter of interest. Finally, in most complex situations EL admits simplified versions, obtained by plugging-in appropriate estimates of parameters, often nuisance parameters. These versions benefit from the reduced computational burden but at the expense of the shape of the approximant distribution, which often changes from the standard χ^2 distribution to a scaled χ^2 distribution (see Hjort et al.,[2] and Adimari and Guolo[3]). In the following, we will call one such version "estimated empirical likelihood."

Nowadays, EL methods have a wide range of applications in various research fields (see, Lazar[4] and Liu and Zhao,[5] for recent reviews), including the receiver operating characteristic (ROC) analysis for a three-class problem, which is commonly used to evaluate the ability of a diagnostic test (or biomarker) to distinguish three ordinal classes (e.g. benign, stage 1, stage 2) of a disease.[6]

Let Y be a diagnostic test result often measured on a continuous scale, and let Y_1, Y_2, Y_3 be the test result for subjects in classes 1, 2, and 3, respectively. Without loss of generality, we assume that higher values of test results are associated with higher severity of the disease. This assumption can be formalized by the simple ordering of the means, i.e. $\mu_1 < \mu_2 < \mu_3$, here μ_j is the mean of Y_j $(j = 1, 2, 3)$. Given a pair of thresholds (t_1, t_2), with $t_1 < t_2$, the true class fractions (TCFs) are defined as[7]

$$\begin{aligned} \theta_1 &\equiv \mathrm{TCF}_1(t_1) = \Pr(Y_1 \le t_1) = F_1(t_1) \\ \theta_2 &\equiv \mathrm{TCF}_2(t_1, t_2) = \Pr(t_1 < Y_2 \le t_2) = F_2(t_2) - F_2(t_1) \\ \theta_3 &\equiv \mathrm{TCF}_3(t_2) = \Pr(Y_3 > t_2) = 1 - F_3(t_2) \end{aligned} \tag{1}$$

where $F_1(\cdot)$, $F_2(\cdot)$ and $F_3(\cdot)$ are the cumulative distribution functions of Y_1, Y_2 and Y_3, respectively. Then, by plotting $(\mathrm{TCF}_1(t_1), \mathrm{TCF}_2(t_1, t_2), \mathrm{TCF}_3(t_2))$ in a unit cube over all possible values of thresholds $t_1 < t_2$, one obtain the ROC surface for the test Y.[7] The volume under the ROC surface (VUS), defined as

$$\gamma = \Pr(Y_1 < Y_2 < Y_3), \tag{2}$$

is usually considered a summary measure of the diagnostic accuracy of the test. The values of VUS vary from $1/6$ to 1, ranging from the chance to perfect diagnostic tests.[7]

Several parametric, semi-parametric, or kernel-based approaches have been developed to estimate the ROC surface; we cite here, among others, papers by Nakas and Yiannoutsos,[7] Xiong et al.,[8] Li and Zhou[9]; and Kang and Tian.[10] The asymptotic normality of the proposed estimators can be used to construct confidence regions at a fixed point of the ROC surface, i.e. for a tern of TCFs, $(\theta_1, \theta_2, \theta_3)$, at a given pair of thresholds (t_1, t_2). To the best of our knowledge, no EL methods are available to attain this last goal.

As for the construction of confidence intervals for the VUS, parametric or nonparametric approaches have been developed; for instance, by Nakas,[6] Xiong et al.,[8] and Guangming et al.[11] Apart from the "hybrid" approach proposed by Guangming et al.[11] which uses jackknife EL,[12] only a "proper" (estimated) EL method has been proposed by Wan,[13] based on the so-called placement values. The author defined an EL pivot and proved the appropriateness of an approximant scaled χ^2 distribution, with an unknown scale constant. Such constant results in a ratio of variances, and can be estimated through a ratio of functions of U-statistics.

Within the three-class ROC analysis framework, the problem of making inferences about a TCF, given the remaining two, is also of interest. Especially, for medical practitioners can be important to evaluate the accuracy of a diagnostic test to distinguish the second class of disease, i.e. to evaluate the probability, θ_2, with which the test correctly classifies a subject at the second stage of disease (often called early stage), when are fixed the values for the true class fractions at first and third classes, θ_1 and θ_3, i.e. when the test simultaneously ensures certain values for the correct probabilities of classification for the first and third stages.

In the last decade, methods have been developed to address the problem of building confidence intervals for θ_2, given θ_1 and θ_3. Dong et al.[14] proposed to use a generalized inference approach for interval estimation of θ_2, under normality assumption or after Box-Cox transformation in non-normal cases. The authors also developed some nonparametric bootstrap-based approaches, referred to as BTP and BTII. Dong and Tian,[15] instead, developed two estimated EL-based methods, ELP and ELB. The authors defined an estimated profile empirical log-likelihood ratio for θ_2 and proved that it approximately follows a scaled χ^2 distribution, with an unknown scale constant which is still a ratio of variances. To estimate the unknown scale constant, the authors proposed, as alternatives, a method involving kernel density estimators, and a bootstrap-based procedure. A different profile empirical log-likelihood for θ_2, and an adjusted version, have been proposed by Rahman and Zhao,[16] referred to as PEL and AEL. They have the advantage of having the standard χ^2 distribution as approximating distribution, but they are more complicated to compute than the competitors mentioned above. Finally, Hai et al.[17] proposed to obtain a confidence interval for θ_2 (given θ_1 and θ_3) by using an estimated EL pivot based on an estimated version of the so-called influence function of an estimator for θ_2. The proposed pivot has a standard χ^2 asymptotic distribution, but the estimation of the influence function involves kernel density estimation.

In summary, the above-mentioned EL techniques have different genesis and solve the problems of building approximate confidence intervals for the TCF at the early stage of disease and for the VUS of a diagnostic test. No EL methods to construct confidence regions for the tern $(\theta_1, \theta_2, \theta_3)$ on the ROC surface, at a fixed pair of thresholds (t_1, t_2), are present in the literature. Moreover, to the best of our knowledge, there are no methods in the literature for constructing confidence regions for a pair of TCFs, the remaining other fixed. In practice, there may be situations in which, for instance, the decision to treat a patient because erroneously positive to the response of a diagnostic test could have high costs (personal or for the health system). In such situations, the researcher may want to require the test to have some fixed level for TCF_1 (the first class typically considers the absence of disease or its harmless level) and study its ability to discriminate between the second and third classes of disease.

This paper aims to propose a unified EL approach that allows to solve the four types of problems considered so far. Our proposal provides EL techniques that are generally easier to use when compared to competing ones. Moreover, the results of an extensive simulation study indicate that our techniques are generally at least as accurate as others (not just those based on EL), and sometimes more accurate in terms of real coverage in finite samples.

The paper is organized as follows. In Section 2, the proposals are laid out in detail, supported by theoretical results. Section 3 presents extensive simulation studies to assess our proposals' performance in finite samples, and compare them with existing methods when possible. Section 4 illustrates our methods in evaluating the ability of some gene expressions to distinguish ductal carcinomas in situ (DCIS) from noncancerous (NC) and invasive breast cancers (IBCs). Finally, a few concluding remarks can be found in Section 5, along with some considerations about various directions for future research.

2 The proposal

2.1 Three-sample EL and confidence regions for triplets of TCFs

Let $\{y_{d1}, \dots, y_{dn_d}\}$ be a random sample from Y_d, the test result for n_d patients from the dth class, $d = 1, 2, 3$. For a fixed t, let $\hat{F}_d(t) = \frac{1}{n_d}\sum_{i=1}^{n_d} \mathrm{I}(y_{di} \le t)$, be the empirical distribution function based on the sample $\{y_{d1}, \dots, y_{dn_d}\}$, and $\hat{P}(t_1, t_2) = \hat{F}_2(t_2) - \hat{F}_2(t_1)$ for fixed t_1 and t_2 $(t_1 < t_2)$, where $\mathrm{I}(\cdot)$ is the indicator function. Let $\boldsymbol{p}_d = (p_{d1}, \dots, p_{dn_d})$ denote a probability vector, with $p_{di} > 0$, parameterizing multinomial distributions on the data points $\{y_{d1}, \dots, y_{dn_d}\}$, with $d = 1, 2, 3$. The EL for $(\theta_1, \theta_2, \theta_3)$, at fixed $t_1 < t_2$, can be defined as

$$L(\theta_1, \theta_2, \theta_3; t_1, t_2) = \sup_{\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3} \prod_{i=1}^{n_1} p_{1i} \prod_{j=1}^{n_2} p_{2j} \prod_{k=1}^{n_3} p_{3k} \tag{3}$$

subject to the following constraints: $p_{1i} > 0, p_{2j} > 0, p_{3k} > 0$, $\sum_{i=1}^{n_1} p_{1i} = 1$, $\sum_{j=1}^{n_2} p_{2j} = 1$, $\sum_{k=1}^{n_3} p_{3k} = 1$,

$$\sum_{i=1}^{n_1} p_{1i}\mathrm{I}(y_{1i} \le t_1) = \theta_1, \quad \sum_{j=1}^{n_2} p_{2j}\mathrm{I}(t_1 < y_{2j} \le t_2) = \theta_2, \quad \sum_{k=1}^{n_3} p_{3k}\mathrm{I}(y_{3k} \le t_2) = 1 - \theta_3.$$

Under the constraints $\sum_{i=1}^{n_1} p_{1i} = 1$, $\sum_{j=1}^{n_2} p_{2j} = 1$ and $\sum_{k=1}^{n_3} p_{3k} = 1$, the EL $L(\theta_1, \theta_2, \theta_3; t_1, t_2)$ in (3) will reach its maximum $n_1^{n_1} n_2^{n_2} n_3^{n_3}$, at $p_{1i} = 1/n_1$, $p_{2j} = 1/n_2$ and $p_{3k} = 1/n_3$ for all i, j, k. Thus, the empirical log-likelihood ratio for $(\theta_1, \theta_2, \theta_3)$ is defined as

$$\ell(\theta_1, \theta_2, \theta_3; t_1, t_2) = \sup_{\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3} \left\{ -2\sum_{i=1}^{n_1} \log\left(n_1 p_{1i}\right) - 2\sum_{j=1}^{n_2} \log\left(n_2 p_{2j}\right) - 2\sum_{k=1}^{n_3} \log\left(n_3 p_{3k}\right) \right\} \tag{4}$$

subject to the constraints mentioned above. In Section A of the Supplemental Materials, we show that

$$\begin{aligned}\ell(\theta_1,\theta_2,\theta_3,;t_1,t_2) = 2n_1 &\left\{ \widehat{F}_1(t_1)\log\frac{\widehat{F}_1(t_1)}{\theta_1} + \left[1-\widehat{F}_1(t_1)\right]\log\frac{1-\widehat{F}_1(t_1)}{1-\theta_1} \right\} \\ + 2n_2 &\left\{ \widehat{P}(t_1,t_2)\log\frac{\widehat{P}(t_1,t_2)}{\theta_2} + \left[1-\widehat{P}(t_1,t_2)\right]\log\frac{1-\widehat{P}(t_1,t_2)}{1-\theta_2} \right\} \\ + 2n_3 &\left\{ \widehat{F}_3(t_2)\log\frac{\widehat{F}_3(t_2)}{1-\theta_3} + \left[1-\widehat{F}_3(t_2)\right]\log\frac{1-\widehat{F}_3(t_2)}{\theta_3} \right\}\end{aligned} \tag{5}$$

when

$$\begin{cases} t_1 \in \left[y_{1(1)}, y_{1(n_1)}\right) \\ t_1 \text{ or } t_2 \in \left[y_{2(1)}, y_{2(n_2)}\right) \\ t_2 \in \left[y_{3(1)}, y_{3(n_3)}\right) \end{cases}$$

otherwise $\ell(\theta_1,\theta_2,\theta_3;t_1,t_2) = +\infty$. It is worth noting that the empirical estimate $\widehat{P}(t_1,t_2)$ can be zero when there are no observations y_{2j} (with $j = 1,2,\dots,n_2$) laying between (t_1,t_2); in this case $\ell(\theta_1,\theta_2,\theta_3;t_1,t_2)$ is also not well-defined. This can happen when the sample size n_2 is very small, and to avoid this drawback, one could use in (5) continuous versions of $\widehat{F}_d$, as suggested in Adimari.[18]

The next theorem establishes the asymptotic behavior of $\ell(\theta_1,\theta_2,\theta_3;t_1,t_2)$. For a fixed pair of thresholds (t_{10},t_{20}) such that $t_{10} < t_{20}$, we denote as $\theta_{10} = F_1(t_{10})$, $\theta_{20} = F_2(t_{20}) - F_2(t_{10})$ and $\theta_{30} = 1 - F_3(t_{20})$ the true values for the parameters of interest.

Theorem 2.1. *If* $\min\{n_1,n_2,n_3\} \to +\infty$, *then we have*

$$\ell(\theta_{10},\theta_{20},\theta_{30};t_{10},t_{20}) \xrightarrow{d} \chi^2_3 \tag{6}$$

where χ^2_3 *indicates the chi-squared distribution with 3 degrees of freedom.*

The proof can be found in Section B of the Supplemental Materials. Based on the result in Theorem 2.1, we can construct a set

$$\mathcal{R}_\alpha = \left\{(\theta_1,\theta_2,\theta_3) : \ell(\theta_1,\theta_2,\theta_3;t_{10},t_{20}) \le \chi^2_{3,(1-\alpha)}\right\}$$

where $\alpha \in (0,1)$ and $\chi^2_{3,(1-\alpha)}$ is the $(1-\alpha)$th quantile of a χ^2_3 distribution. $\mathcal{R}_\alpha$ is a (nonparametric) confidence region, with nominal coverage probability $1-\alpha$, for a TCFs triplet $(\theta_{10},\theta_{20},\theta_{30})$, at a fixed pair of thresholds (t_{10},t_{20}). This 3D confidence region can be easily produced using the `contour3d()` function of the `R`[19] package `misc3d`.[20] In what follows, we will denote such confidence regions as ELQ3D.

It is worth noting that the nonparametric log-likelihood ratio in (5), coincides with the log-likelihood ratio based on the (independent) binomial distributions of $\widehat{F}_1(t_1)$, $\widehat{P}_2(t_1,t_2)$ and $\widehat{F}_3(t_2)$. This provides a further justification for the result in Theorem 2.1 under the considered weak conditions, which do not even involve the continuity of F_1, F_2 and F_3.

2.2 Confidence intervals for TCF$_2$

In some circumstances, one could have enough information to fix the values for θ_1 and θ_3, or one may want to fix such quantities to desired values. Then, thresholds t_1 and t_2 could be estimated. Let $\widehat{t}_1 = \widehat{F}_1^{-1}(\theta_1) = \inf\{t : \widehat{F}_1(t) \ge \theta_1\}$ and $\widehat{t}_2 = \widehat{F}_3^{-1}(1-\theta_3)$ be estimates of t_1 and t_2, respectively, for fixed θ_1 and θ_3. In this situation, by using the plug-in method, we have an estimated version of the EL statistic $\ell(\theta_1, \theta_2, \theta_3; t_1, t_2)$ as follows:

$$
\begin{aligned}
\ell_*(\theta_2) &= \ell(\theta_1, \theta_2, \theta_3; \hat{t}_1, \hat{t}_2) \\
&= 2n_1 \left\{ \hat{F}_1(\hat{t}_1) \log \frac{\hat{F}_1(\hat{t}_1)}{\theta_1} + \left[1 - \hat{F}_1(\hat{t}_1)\right] \log \frac{1 - \hat{F}_1(\hat{t}_1)}{1 - \theta_1} \right\} \\
&\quad + 2n_2 \left\{ \hat{P}(\hat{t}_1, \hat{t}_2) \log \frac{\hat{P}(\hat{t}_1, \hat{t}_2)}{\theta_2} + \left[1 - \hat{P}(\hat{t}_1, \hat{t}_2)\right] \log \frac{1 - \hat{P}(\hat{t}_1, \hat{t}_2)}{1 - \theta_2} \right\} \\
&\quad + 2n_3 \left\{ \hat{F}_3(\hat{t}_2) \log \frac{\hat{F}_3(\hat{t}_2)}{1 - \theta_3} + \left[1 - \hat{F}_3(\hat{t}_2)\right] \log \frac{1 - \hat{F}_3(\hat{t}_2)}{\theta_3} \right\}
\end{aligned} \tag{7}
$$

if $\hat{t}_1$ or $\hat{t}_2 \in [y_{2(1)}, y_{2(n_2)})$, otherwise $\ell_*(\theta_2) = +\infty$. Substitution of t_1 and t_2 by their estimates impacts on the standard χ^2 approximation, that no longer holds. The following theorem shows that $\ell_*(\theta_{20})$, for fixed $\theta_1 = \theta_{10}$ and $\theta_3 = \theta_{30}$, asymptotically follows a scaled χ^2 distribution, under some smoothness conditions. Again, θ_{20} denotes the true value for the parameter of interest.

Theorem 2.2. *Assume that F_1, F_2 and F_3 have continuous density functions f_1, f_2 and f_3 such that $f_1(t_{10}) > 0$, $f_2(t_{10}) > 0$, $f_2(t_{20}) > 0$ and $f_3(t_{20}) > 0$. Let $t_{10} = F_1^{-1}(\theta_{10})$ and $t_{20} = 1 - F_3^{-1}(\theta_{30})$. If $\min\{n_1, n_2, n_3\} \to +\infty$ and the ratios n_1/n_2, n_3/n_2 have finite non-zero limits, then*

$$
w\ell_*(\theta_{20}) = w\ell(\theta_{10}, \theta_{20}, \theta_{30}; \hat{t}_{10}, \hat{t}_{20}) \xrightarrow{d} \chi_1^2 \tag{8}
$$

where $w > 0$ is a suitable finite parameter.

The proof can be found in Section C of the Supplemental Materials, along with the expression of parameter w. To estimate w, one could use estimators for variances, involving kernel estimation of $f_1(\cdot)$, $f_2(\cdot)$, and $f_3(\cdot)$ (see the proof of Theorem 2.2). However, we observe that, from (8), the quantity $w\mathrm{Med}(\ell_*(\theta_{20}))$ is asymptotically equal to $\mathrm{Med}(\chi_1^2)$ which is $(7/9)^3$; here $\mathrm{Med}(\cdot)$ stands for the median operator. Thus, we propose to estimate w by $\hat{w} = \dfrac{(7/9)^3}{\widehat{\mathrm{Med}}(\ell_*(\theta_{20}))}$, with $\widehat{\mathrm{Med}}(\ell_*(\theta_{20}))$ obtained by the following bootstrap procedure:

1. from observed data $y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}$ and $y_{31}, \ldots, y_{3n_3}$, obtain the estimates $\hat{t}_{10}$, $\hat{t}_{20}$ and $\hat{\theta}_{20} = \hat{F}_2(\hat{t}_{20}) - \hat{F}_2(\hat{t}_{10})$;
2. get B bootstrap samples $\{y_1\}_b$, $\{y_2\}_b$ and $\{y_3\}_b$, for $b = 1, \ldots, B$, of sizes n_1, n_2 and n_3, respectively;
3. from the bth tern of bootstrap samples, compute the estimates $\hat{t}_{10b}$ and $\hat{t}_{20b}$, and then, $\ell_{*b}(\hat{\theta}_{20}) = \ell(\theta_{10}, \hat{\theta}_{20}, \theta_{30}; \hat{t}_{10b}, \hat{t}_{20b})$;
4. get the estimate $\widehat{\mathrm{Med}}(\ell_*(\theta_{20}))$ as the sample median from the values $\ell_{*b}(\hat{\theta}_{20})$, $b = 1, \ldots, B$.

In the above-given procedure, only those bootstrap samples are processed whose sample averages respect the fixed ordering of the population means ($\mu_1 < \mu_2 < \mu_3$). To improve accuracy in the estimation of w, in step 4 the continuous version of $\hat{F}_j$[18] is employed when computing, from each bootstrap sample, $\ell_{*b}(\hat{\theta}_{20})$. The choice to use bootstrap to estimate the median of $\ell_*(\theta_{20})$ is dictated by the simplicity of this approach with respect, for instance, to the bootstrap calibration, and by the fact that resorting to the median avoids suitable *ad hoc* measures to treat situations in which $\ell_{*b}(\hat{\theta}_{20})$ assumes not finite value.

A confidence interval for θ_{20}, for fixed $\theta_1 = \theta_{10}$ and $\theta_3 = \theta_{30}$, with nominal coverage $1 - \alpha$, is therefore obtained as

$$
\mathcal{R}_{2,\alpha}^* = \left\{ \theta_2 : \hat{w}\ell_*(\theta_2) \leq \chi^2_{1,(1-\alpha)} \right\}
$$

where $\alpha \in (0, 1)$ and $\chi^2_{1,(1-\alpha)}$ is the $(1 - \alpha)$th quantile of a χ_1^2 distribution. The interval $\mathcal{R}_{2,\alpha}^*$ will be denoted as ELQB.

As noted by one reviewer, the first and third addend in (7) are asymptotically negligible. One might therefore define $\ell_*(\theta_2)$ to be equal to the second addend only. Although, at large sample sizes, the two alternative definitions are equivalent in terms of the agreement between their distribution and the asymptotic counterpart, we observe that at low sample sizes, where the impact of individual terms can be more pronounced, the presence of these negligible terms helps achieve a better agreement between the two distributions. A similar effect can be observed in one of the quantities which will be later introduced, namely (13), whose first addend is also asymptotically negligible. For this reason, we prefer to keep all terms in the definitions introduced in the paper. This, in our opinion, also helps the presentation and understanding of the contents of the paper itself.

2.3 Confidence intervals for the VUS

Observe that, in (5), each of the three addenda is an EL pivot (with approximant distribution χ_1^2) for inference about an unknown proportion or probability, estimated by its empirical counterpart; in a sample of independent and identically distributed observations, such empirical counterpart, multiplied by the sample size, is the realization of a binomial random variable.

Let γ be the unknown VUS for a diagnostic test. As γ is a probability, our idea is to use the quantity

$$\ell(\gamma) = 2n\left\{\widehat{\gamma}\log\frac{\widehat{\gamma}}{\gamma} + (1-\widehat{\gamma})\log\frac{1-\widehat{\gamma}}{1-\gamma}\right\} \tag{9}$$

to obtain a pivot for interval estimation of the VUS. In (9), $n = n_1 + n_2 + n_3$ and

$$\widehat{\gamma} = \widehat{\Pr}(Y_1 < Y_2 < Y_3) = \frac{1}{n_1 n_2 n_3}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} \mathrm{I}\left(y_{1i} < y_{2j} < y_{3k}\right) \tag{10}$$

is an estimate, given by an unbiased nonparametric estimator.[7] Assuming $\gamma \in (0,1)$, $\ell(\gamma)$ in (9) is well-defined if $\widehat{\gamma}$ is different from 1 (or 0). However, the estimate $\widehat{\gamma}$ is a three-sample U-statistic, and its presence in (9) affects the asymptotic behavior of the quantity itself. We prove the following theorem.

Theorem 2.3. *Let γ_0 be the true value of $\gamma \in (0,1)$. When $\min\{n_1, n_2, n_3\} \to +\infty$ and $n_d/n \to \rho_d$, with $0 < \rho_d < 1$, $d = 1, 2, 3$, then*

$$w\ell(\gamma_0) \xrightarrow{d} \chi_1^2 \tag{11}$$

where $w > 0$ is a suitable finite parameter.

The proof can be found in Section D of the Supplemental Materials, along with the expression of parameter w. One could estimate the scale parameter w as $\dfrac{\widehat{\gamma}(1-\widehat{\gamma})}{n\,\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\gamma})}$, where $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\gamma})$ is the estimated variance of $\widehat{\gamma}$.[21] However, we follow the same idea as in Section 2.2, and propose to estimate w by $\widehat{w} = \dfrac{(7/9)^3}{\widehat{\mathrm{Med}}(\ell(\gamma_0))}$, with the estimate $\widehat{\mathrm{Med}}(\ell(\gamma_0))$ obtained by the following bootstrap procedure:

(1) from the observed data $y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}$ and $y_{31}, \ldots, y_{3n_3}$, obtain the estimate $\widehat{\gamma}$ by (10);
(2) get B tern of bootstrap samples $\{y_1\}_b$, $\{y_2\}_b$ and $\{y_3\}_b$, for $b = 1, \ldots, B$, of sizes n_1, n_2 and n_3, respectively;
(3) from the bth tern of bootstrap samples, compute the estimate $\widehat{\gamma}_b$, and then, $\ell_b(\widehat{\gamma})$;
(4) get the estimate $\widehat{\mathrm{Med}}(\ell(\gamma_0))$ as the sample median from the values $\ell_b(\widehat{\gamma})$, $b = 1, \ldots, B$.

Again, only those bootstrap samples whose sample averages respect the fixed ordering of the population means ($\mu_1 < \mu_2 < \mu_3$) are processed. Moreover, to avoid technical problems, we set $\widehat{\gamma}_b = (n_1 n_2 n_3)/(n_1 n_2 n_3 + 0.5)$, when it happens that $\widehat{\gamma}_b = 1$. The confidence interval for γ_0 is therefore obtained as

$$\mathcal{R}_{\gamma,\alpha} = \left\{\gamma : \widehat{w}\ell(\gamma) \le \chi^2_{1,(1-\alpha)}\right\}$$

where $\alpha \in (0,1)$ and $\chi^2_{1,(1-\alpha)}$ is the $(1-\alpha)$th quantile of a χ_1^2 distribution. The interval $\mathcal{R}_{\gamma,\alpha}$ will denoted as ELQB.

The technique proposed to build confidence intervals for the VUS can be easily extended to the case where ties are present in the samples. Our approach here does not require the distribution functions F_1, F_2, and F_3 to be continuous.

When ties are present in the data, it is sufficient to use in (9) the appropriate estimate of γ, i.e.

$$\widehat{\gamma} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \Big\{ \mathrm{I}\left(y_{1i} < y_{2j} < y_{3k}\right) + \frac{1}{2}\mathrm{I}\left(y_{1i} = y_{2j} < y_{3k}\right) + \frac{1}{2}\mathrm{I}\left(y_{1i} < y_{2j} = y_{3k}\right) + \frac{1}{6}\mathrm{I}\left(y_{1i} = y_{2j} = y_{3k}\right) \Big\}. \tag{12}$$

2.4 Confidence regions for the pair (TCF_2,TCF_3)

Suppose now that the researcher fixes the value θ_1 for TCF_1. Let $\widehat{t}_1 = \widehat{F}_1^{-1}(\theta_1)$. By using the plugin method, again, we have an estimated version of EL statistic $\ell(\theta_1, \theta_2, \theta_3; t_1, t_2)$ as follows:

$$\begin{aligned}
\ell_{**}(\theta_2, \theta_3; t_2) &= \ell(\theta_1, \theta_2, \theta_3; \widehat{t}_1, t_2) \\
&= 2n_1 \left\{ \widehat{F}_1(\widehat{t}_1) \log \frac{\widehat{F}_1(\widehat{t}_1)}{\theta_1} + \left[1 - \widehat{F}_1(\widehat{t}_1)\right] \log \frac{1 - \widehat{F}_1(\widehat{t}_1)}{1 - \theta_1} \right\} \\
&\quad + 2n_2 \left\{ \widehat{P}(\widehat{t}_1, t_2) \log \frac{\widehat{P}(\widehat{t}_1, t_2)}{\theta_2} + \left[1 - \widehat{P}(\widehat{t}_1, t_2)\right] \log \frac{1 - \widehat{P}(\widehat{t}_1, t_2)}{1 - \theta_2} \right\} \\
&\quad + 2n_3 \left\{ \widehat{F}_3(t_2) \log \frac{\widehat{F}_3(t_2)}{1 - \theta_3} + \left[1 - \widehat{F}_3(t_2)\right] \log \frac{1 - \widehat{F}_3(t_2)}{\theta_3} \right\}
\end{aligned} \tag{13}$$

given θ_1, if

$$\begin{cases} \widehat{t}_1 \text{ or } t_2 \in \left[y_{2(1)}, y_{2(n_2)}\right) \\ t_2 \in \left[y_{3(1)}, y_{3(n_3)}\right) \end{cases}$$

otherwise $\ell_{**}(\theta_2, \theta_3; t_2) = +\infty$. The following theorem shows how to obtain from $\ell_{**}(\theta_2, \theta_3; t_2)$ regions for the true pair $(\theta_{20}, \theta_{30})$ at a fixed value t_{20} for the second threshold, given $\theta_1 = \theta_{10}$.

Theorem 2.4. *Assume that F_1 and F_2 have continuous density functions f_1 and f_2, such that $f_1(t_{10}) > 0$ and $f_2(t_{10}) > 0$. Let $t_{10} = F_1^{-1}(\theta_{10})$. If $\min\{n_1, n_2, n_3\} \to +\infty$ and the ratio n_1/n_2 has finite non-zero limit, then*

$$\ell_{**}(\theta_{20}, \theta_{30}; t_{20}) = \ell(\theta_{10}, \theta_{20}, \theta_{30}; \widehat{t}_{10}, t_{20}) \xrightarrow{d} wU_1 + U_2 \tag{14}$$

where $w > 0$ is a suitable finite parameter, U_1 and U_2 are independent χ_1^2 random variables.

The proof can be found in Section E of the Supplemental Materials, along with the expression of parameter w. From the results in the proof, to estimate w, we propose to use $\widehat{w} = \dfrac{\widehat{\mathrm{Med}}(\ell_2(\theta_{20}; \widehat{t}_{10}, t_{20}))}{(7/9)^3}$, with

$$\ell_2(\theta_{20}; \widehat{t}_{10}, t_{20}) = 2n_2 \left\{ \widehat{P}(\widehat{t}_{10}, t_{20}) \log \frac{\widehat{P}(\widehat{t}_{10}, t_{20})}{\theta_{20}} + \left[1 - \widehat{P}(\widehat{t}_{10}, t_{20})\right] \log \frac{1 - \widehat{P}(\widehat{t}_{10}, t_{20})}{1 - \theta_{20}} \right\}$$

and $\widehat{\mathrm{Med}}(\ell_2(\theta_{20}; \widehat{t}_{10}, t_{20}))$ obtained by the following bootstrap procedure:

(1) from observed data $y_{11}, \ldots, y_{1n_1}$ and $y_{21}, \ldots, y_{2n_2}$ obtain the estimates $\widehat{t}_{10} = \widehat{F}_1^{-1}(\theta_{10})$ and $\widehat{\theta}_{20} = \widehat{F}_2(t_{20}) - \widehat{F}_2(\widehat{t}_{10})$;
(2) get B bootstrap samples $\{y_1\}_b$ and $\{y_2\}_b$ for $b = 1, \ldots, B$, of sizes n_1 and n_2, respectively;
(3) from the bth pair of the bootstrap sample, compute the estimate $\widehat{t}_{10b}$, and then, $\ell_{2b}(\widehat{\theta}_{20}; \widehat{t}_{10b}, t_{20})$;
(4) get the estimate $\widehat{\mathrm{Med}}(\ell_2(\theta_{20}; \widehat{t}_{10}, t_{20}))$ as the sample median from the values $\ell_{2b}(\widehat{\theta}_{20}; \widehat{t}_{10b}, t_{20})$, $b = 1, \ldots, B$.

Table 1. Scenarios for the simulation study.

Scenario	Y_1	Y_2	Y_3	t_{10}	t_{20}	θ_{10}	θ_{20}	θ_{30}	γ_0
1	$\mathcal{N}(0,1)$	$\mathcal{N}(2.5,1.1^2)$	$\mathcal{N}(3.69,1.2^2)$	0.842	2.680	0.8	0.5	0.8	0.772
2	$\mathcal{N}(0,1)$	$\mathcal{N}(3.5,1.1^2)$	$\mathcal{N}(5.5,1.2^2)$	0.842	4.490	0.8	0.8	0.8	0.881
3	$\mathcal{N}(0,1)$	$\mathcal{N}(4,1.2^2)$	$\mathcal{N}(8.189,2^2)$	1.282	5.626	0.9	0.9	0.9	0.959
4	$\mathcal{G}(6,12)$	$\mathcal{LN}(1.5,0.5)$	$\mathcal{W}(4,6.6)$	0.659	4.536	0.8	0.5	0.8	0.669
5	$\mathcal{G}(6,12)$	$\mathcal{LN}(1.5,0.5)$	$\mathcal{W}(4,10)$	0.659	6.873	0.8	0.8	0.8	0.868
6	$\mathcal{G}(6,12)$	$\mathcal{LN}(1.5,0.5)$	$\mathcal{W}(4,12.4)$	0.659	8.523	0.8	0.9	0.8	0.927
7	$\mathcal{B}(1,6)$	$\mathcal{B}(6,6)$	$\mathcal{B}(9.6,6)$	0.235	0.513	0.8	0.5	0.8	0.698
8	$\mathcal{B}(1,6)$	$\mathcal{B}(9,6)$	$\mathcal{B}(20.4,6)$	0.235	0.707	0.8	0.8	0.8	0.869
9	$\mathcal{B}(1,6)$	$\mathcal{B}(6,6)$	$\mathcal{B}(20.4,6)$	0.235	0.707	0.8	0.9	0.8	0.917
10	$0.5N(-1,1)+$ $0.5N(2,1)$	$0.5N(1,1)+$ $0.5N(4,1.5)$	$0.5N(3,1.5)+$ $0.5N(6,1)$	0.5	4.5	0.5	0.674	0.522	0.544

Here, $\mathcal{N}$, $\mathcal{LN}$, $\mathcal{G}$, $\mathcal{B}$ and $\mathcal{W}$ indicate normal, log-normal, gamma, beta and Weibull distributions; θ_{10}, θ_{20} and θ_{30} are true values of TCFs; and γ_0 is the true value of VUS. TCF: true class fraction; VUS: volume under the receiver operating characteristic surface.

Note that the bootstrap procedure of Section 2.2 apply also here. Then, fixed $\theta_1 = \theta_{10}$, a confidence region for the pair $(\theta_{20}, \theta_{30})$, with nominal coverage $1-\alpha$, is obtained as

$$\mathcal{R}^*_{23,\alpha} = \left\{(\theta_2, \theta_3) : \ell_{**}(\theta_2, \theta_3; t_{20}) \leq \widehat{c}_\alpha\right\}$$

where $\alpha \in (0, 1)$ and $\widehat{c}_\alpha$ is the sample quantile of order $(1-\alpha)$, from 1000 Monte Carlo values generated as $\widehat{w}U_1 + U_2$, where U_1 and U_2 are independent χ^2_1 random variables. The interval $\mathcal{R}^*_{23,\alpha}$ will be denoted as ELQB. Of course, the approach proposed here can also be used to construct confidence regions for a different pair of TCFs, such as $(\mathrm{TCF}_1, \mathrm{TCF}_3)$, fixed $\theta_2 = \theta_{20}$.

3 Simulation study

3.1 Simulation set-up

To investigate the finite sample behavior of our proposed EL techniques, based on $\ell(\theta_1, \theta_2, \theta_3; t_1, t_2)$ in (5), $\ell_*(\theta_2)$ in (7), $l(\gamma)$ in (9) and $l_{**}(\theta_2, \theta_3; t_2)$ in (13), we conducted a large simulation study. In particular, we evaluated the coverage probability of 3D confidence regions for the triplet of TCFs $(\theta_{10}, \theta_{20}, \theta_{30})$, at fixed thresholds (t_{10}, t_{20}), of confidence intervals for TCF_2, θ_{20}, at fixed TCF_1 and TCF_3, of confidence intervals for the VUS and confidence regions for the pair $(\mathrm{TCF}_2, \mathrm{TCF}_3)$, $(\theta_{20}, \theta_{30})$, at fixed TCF_1. We also compared our proposals with competitors, when present in the literature.

In the simulation experiments, we considered ten scenarios, listed in Table 1. In particular, the first three scenarios refer to the tri-normal setting; scenarios 4 to 6 refer to the case of mixed distributions (gamma, log-normal, and Weibull distributions); scenarios 7 to 9 consider a tri-beta setting, where the biomarkers' values are bounded in (0, 1); the last scenario refers to a setting that considers mixture distributions.

3.2 Results: Confidence regions for TCFs

The performance of our proposed EL confidence region ELQ3D is evaluated at three different levels of nominal coverage $1-\alpha$, i.e. 0.90, 0.95, and 0.99. Under each scenario in Table 1, 10,000 random samples are generated. The sample sizes (n_1, n_2, n_3) are set as (30, 30, 30), (50, 50, 50), and (100, 100, 100). The simulation results are presented in Table 2.

As one can see, the empirical coverages are close to the nominal ones in almost all the considered settings. As expected, our EL confidence region ELQ3D needs a larger sample size when the true TCFs are close to 1.

3.3 Results: Confidence intervals for TCF_2, at fixed θ_{10} and θ_{30}

Here, we compare the performance of our proposed methods ELQB in Section (2.2) with the existing nonparametric approaches, i.e. ELP, ELB, and BTII,[15] IF,[17] PEL and AEL,[16] through scenarios in Table 1. Under each scenario, we fixed the values θ_{10} and θ_{30}, and generated 5000 random samples. The sample sizes (n_1, n_2, n_3) are set at (30, 30, 30), (50, 30, 30), (50, 50, 50), (100, 50, 50), (100, 100, 50), and (100, 100, 100). For, the ELB and BTII methods, we consider 500 bootstrap samples, whereas, for our methods, we use $B = 200$. Simulation results are reported in Tables 3 to 6. Although

Table 2. Monte Carlo coverages for the ELQ3D confidence regions for $(\theta_{10}, \theta_{20}, \theta_{30})$, at fixed values for t_1 and t_2, for each scenario in Table 1.

Scenario	θ_{10}	θ_{20}	θ_{30}	$n_1 = n_2 = n_3$	Nominal level		
					0.90	0.95	0.99
1	0.8	0.5	0.8	30	0.892	0.944	0.988
				50	0.902	0.950	0.991
				100	0.896	0.943	0.990
2	0.8	0.8	0.8	30	0.890	0.939	0.988
				50	0.896	0.945	0.980
				100	0.898	0.947	0.989
3	0.9	0.9	0.9	30	0.833	0.859	0.874
				50	0.895	0.945	0.986
				100	0.895	0.949	0.990
4	0.8	0.5	0.8	30	0.903	0.951	0.989
				50	0.904	0.950	0.991
				100	0.902	0.950	0.990
5	0.8	0.8	0.8	30	0.894	0.943	0.989
				50	0.897	0.947	0.990
				100	0.896	0.944	0.987
6	0.8	0.9	0.8	30	0.866	0.910	0.946
				50	0.895	0.945	0.986
				100	0.895	0.949	0.990
7	0.8	0.5	0.8	30	0.893	0.947	0.988
				50	0.897	0.946	0.990
				100	0.897	0.947	0.990
8	0.8	0.8	0.8	30	0.894	0.943	0.988
				50	0.894	0.947	0.988
				100	0.898	0.947	0.990
9	0.8	0.9	0.8	30	0.867	0.911	0.948
				50	0.897	0.947	0.984
				100	0.899	0.949	0.989
10	0.5	0.674	0.522	30	0.891	0.945	0.989
				50	0.898	0.948	0.988
				100	0.900	0.949	0.990

BTII is not an EL-based method, it is considered here as a reference, due to its nonparametric nature and its relative ease of use.

Overall, results in Tables 3 to 6 (and Table S1, Section F of the Supplemental Materials) seem to show that the proposed method, ELQB, competes with the best contestants (ELB, BTII), arriving at outperforming in some scenarios, in particular in scenario 10, with mixture models (see Table 6). In general, the approaches IF, AEL, and PEL behave poorly. The IF approach requires greater sample sizes (than that of other approaches) in some scenarios (3, Table 3 and 6, Table 4). The AEL and PEL approaches seem, sometimes, not to yield consistent results even when n grows (see scenario 3 in Table S1, Section F of the Supplemental Materials). Moreover, the three methods seem not to be able to cope with mixture models (Table 6).

Of course, fully nonparametric approaches, such as those here considered, typically perform well only with large sample sizes when true TCFs values are close to 1. In such situations, a sample size of at least greater than 100 for each class is required.

3.4 Results: Confidence intervals for the VUS

We examine the performance of our proposed EL-based method ELQB in Section (2.3), for constructing confidence intervals for the VUS. We also compare our method to the existing approaches, ELU[13] and JEL,[11] through scenarios in Table 1. Under each scenario, we generated 5000 random samples. The sample sizes (n_1, n_2, n_3) are set at (15, 15, 15), (30, 30, 30), (50, 30, 30), (50, 50, 50), (75, 75, 75), and (100, 100, 100). Simulation results are reported in Tables 7 to 10. We also considered some additional scenarios, where the true VUS value is around 0.45 to 0.55; corresponding simulation results are given in Table S2, Section G, of the Supplemental Materials.

Table 3. Monte Carlo coverages for the proposed ELQB confidence intervals for θ_{20}, at fixed θ_{10} and θ_{30}.

Sample size	$1-\alpha$	ELQB	ELP	ELB	IF	PEL	AEL	BTII
Scenario 1: $\mathcal{N}(0,1)$, $\mathcal{N}(2.5,1.1^2)$, $\mathcal{N}(3.69,1.2^2)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.5$								
(30, 30, 30)	0.90	0.900	0.894	0.919	0.874	0.912	0.927	0.909
	0.95	0.949	0.949	0.963	0.931	0.956	0.963	0.951
	0.99	0.988	0.990	0.993	0.982	0.985	0.992	0.985
(50, 30, 30)	0.90	0.894	0.893	0.921	0.873	0.907	0.923	0.917
	0.95	0.942	0.950	0.964	0.932	0.954	0.962	0.953
	0.99	0.985	0.991	0.995	0.983	0.983	0.991	0.985
(50, 50, 50)	0.90	0.901	0.896	0.914	0.889	0.905	0.912	0.918
	0.95	0.947	0.951	0.960	0.940	0.943	0.949	0.954
	0.99	0.985	0.990	0.993	0.985	0.977	0.983	0.989
(100, 50, 50)	0.90	0.892	0.894	0.915	0.885	0.905	0.912	0.920
	0.95	0.942	0.948	0.957	0.936	0.939	0.949	0.956
	0.99	0.982	0.990	0.992	0.983	0.975	0.984	0.987
(100, 100, 50)	0.90	0.877	0.890	0.904	0.885	0.890	0.899	0.913
	0.95	0.928	0.944	0.949	0.939	0.937	0.941	0.951
	0.99	0.977	0.985	0.988	0.982	0.974	0.978	0.984
(100, 100, 100)	0.90	0.897	0.897	0.913	0.892	0.893	0.902	0.921
	0.95	0.944	0.949	0.956	0.943	0.933	0.936	0.961
	0.99	0.986	0.992	0.991	0.990	0.975	0.977	0.990
Scenario 2: $\mathcal{N}(0,1)$, $\mathcal{N}(3.5,1.1^2)$, $\mathcal{N}(5.5,1.2^2)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.8$								
(30, 30, 30)	0.90	0.937	0.936	0.946	0.878	0.919	0.938	0.927
	0.95	0.973	0.973	0.978	0.932	0.951	0.957	0.964
	0.99	0.986	0.999	0.997	0.951	0.986	0.987	0.991
(50, 30, 30)	0.90	0.917	0.932	0.940	0.867	0.908	0.931	0.920
	0.95	0.961	0.975	0.974	0.928	0.947	0.951	0.955
	0.99	0.980	0.997	0.995	0.948	0.986	0.988	0.987
(50, 50, 50)	0.90	0.901	0.911	0.923	0.886	0.916	0.925	0.922
	0.95	0.949	0.960	0.962	0.933	0.951	0.963	0.960
	0.99	0.989	0.992	0.991	0.981	0.984	0.988	0.992
(100, 50, 50)	0.90	0.899	0.909	0.923	0.887	0.910	0.921	0.925
	0.95	0.950	0.961	0.967	0.938	0.950	0.962	0.963
	0.99	0.988	0.993	0.992	0.981	0.983	0.988	0.991
(100, 100, 50)	0.90	0.861	0.899	0.905	0.877	0.882	0.889	0.910
	0.95	0.921	0.953	0.949	0.934	0.934	0.939	0.948
	0.99	0.978	0.991	0.991	0.981	0.974	0.978	0.988
(100, 100, 100)	0.90	0.883	0.895	0.907	0.879	0.888	0.892	0.911
	0.95	0.936	0.949	0.953	0.933	0.939	0.941	0.954
	0.99	0.982	0.991	0.990	0.980	0.974	0.976	0.988
Scenario 3: $\mathcal{N}(0,1)$, $\mathcal{N}(4,1.2^2)$, $\mathcal{N}(8.189,2^2)$, $\theta_{10}=\theta_{30}=0.9$, $\theta_{20}=0.9$								
(30, 30, 30)	0.90	0.826	0.971	0.958	0.806	0.904	0.922	0.951
	0.95	0.839	0.987	0.978	0.820	0.952	0.956	0.976
	0.99	0.849	0.997	0.995	0.826	0.977	0.977	0.995
(50, 30, 30)	0.90	0.821	0.971	0.958	0.803	0.896	0.918	0.949
	0.95	0.834	0.986	0.977	0.816	0.951	0.954	0.972
	0.99	0.845	0.998	0.996	0.824	0.970	0.974	0.993
(50, 50, 50)	0.90	0.921	0.945	0.949	0.864	0.903	0.904	0.945
	0.95	0.937	0.980	0.976	0.888	0.944	0.947	0.974
	0.99	0.952	0.998	0.997	0.896	0.982	0.984	0.995
(100, 50, 50)	0.90	0.915	0.941	0.948	0.859	0.893	0.900	0.938
	0.95	0.932	0.978	0.974	0.882	0.936	0.938	0.971
	0.99	0.949	0.997	0.993	0.891	0.979	0.982	0.991
(100, 100, 50)	0.90	0.875	0.905	0.929	0.849	0.852	0.860	0.930
	0.95	0.933	0.956	0.964	0.897	0.909	0.928	0.960
	0.99	0.981	0.991	0.991	0.936	0.970	0.972	0.989
(100, 100, 100)	0.90	0.893	0.917	0.934	0.882	0.895	0.902	0.935
	0.95	0.943	0.961	0.966	0.929	0.941	0.943	0.965
	0.99	0.986	0.993	0.993	0.965	0.975	0.976	0.993

Normal distributions. ELP, ELB, IF, PEL, AEL and BTII are competitor approaches.

Table 4. Monte Carlo coverages for the proposed ELQB confidence intervals for θ_{20}, at fixed θ_{10} and θ_{30}.

Sample size	$1-\alpha$	ELQB	ELP	ELB	IF	PEL	AEL	BTII
Scenario 4: $\mathcal{G}(6,12)$, $\mathcal{LN}(1.5,0.5)$, $\mathcal{W}(4,6.6)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.5$								
(30, 30, 30)	0.90	0.913	0.899	0.918	0.895	0.922	0.938	0.916
	0.95	0.960	0.948	0.962	0.943	0.966	0.973	0.954
	0.99	0.988	0.987	0.990	0.985	0.991	0.996	0.984
(50, 30, 30)	0.90	0.912	0.898	0.921	0.896	0.927	0.943	0.918
	0.95	0.957	0.945	0.959	0.942	0.968	0.973	0.954
	0.99	0.988	0.989	0.992	0.986	0.990	0.995	0.985
(50, 50, 50)	0.90	0.900	0.883	0.905	0.881	0.913	0.922	0.913
	0.95	0.947	0.939	0.954	0.934	0.953	0.961	0.951
	0.99	0.986	0.985	0.989	0.983	0.984	0.990	0.984
(100, 50, 50)	0.90	0.893	0.887	0.909	0.884	0.915	0.924	0.917
	0.95	0.945	0.941	0.956	0.940	0.952	0.962	0.955
	0.99	0.986	0.985	0.990	0.984	0.984	0.992	0.985
(100, 100, 50)	0.90	0.879	0.877	0.898	0.876	0.902	0.911	0.908
	0.95	0.928	0.933	0.944	0.933	0.943	0.946	0.948
	0.99	0.978	0.984	0.985	0.984	0.979	0.983	0.985
(100, 100, 100)	0.90	0.892	0.888	0.906	0.888	0.898	0.910	0.912
	0.95	0.942	0.940	0.951	0.939	0.941	0.943	0.955
	0.99	0.986	0.988	0.990	0.987	0.983	0.986	0.990
Scenario 5: $\mathcal{G}(6,12)$, $\mathcal{LN}(1.5,0.5)$, $\mathcal{W}(4,10)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.8$								
(30, 30, 30)	0.90	0.939	0.926	0.941	0.907	0.923	0.946	0.931
	0.95	0.979	0.978	0.973	0.947	0.957	0.963	0.963
	0.99	0.993	0.998	0.997	0.956	0.992	0.993	0.991
(50, 30, 30)	0.90	0.931	0.928	0.939	0.901	0.930	0.952	0.928
	0.95	0.973	0.976	0.974	0.948	0.962	0.963	0.964
	0.99	0.992	0.998	0.996	0.959	0.991	0.992	0.991
(50, 50, 50)	0.90	0.912	0.921	0.928	0.904	0.920	0.934	0.931
	0.95	0.959	0.964	0.967	0.951	0.954	0.968	0.964
	0.99	0.993	0.996	0.995	0.988	0.988	0.992	0.993
(100, 50, 50)	0.90	0.904	0.920	0.922	0.902	0.915	0.928	0.922
	0.95	0.948	0.960	0.962	0.951	0.950	0.964	0.959
	0.99	0.991	0.994	0.993	0.984	0.987	0.992	0.991
(100, 100, 50)	0.90	0.897	0.915	0.912	0.905	0.905	0.914	0.923
	0.95	0.944	0.958	0.961	0.948	0.953	0.956	0.962
	0.99	0.986	0.993	0.993	0.987	0.983	0.985	0.990
(100, 100, 100)	0.90	0.902	0.917	0.917	0.909	0.900	0.901	0.924
	0.95	0.955	0.962	0.959	0.957	0.951	0.953	0.963
	0.99	0.991	0.993	0.994	0.990	0.986	0.987	0.993
Scenario 6: $\mathcal{G}(6,12)$, $\mathcal{LN}(1.5,0.5)$, $\mathcal{W}(4,12.4)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.9$								
(30, 30, 30)	0.90	0.891	0.964	0.960	0.804	0.914	0.940	0.918
	0.95	0.908	0.985	0.978	0.822	0.955	0.965	0.952
	0.99	0.919	0.998	0.996	0.841	0.969	0.976	0.982
(50, 30, 30)	0.90	0.903	0.974	0.968	0.827	0.924	0.952	0.933
	0.95	0.917	0.991	0.984	0.837	0.963	0.974	0.960
	0.99	0.927	0.998	0.998	0.852	0.976	0.981	0.989
(50, 50, 50)	0.90	0.948	0.943	0.943	0.898	0.916	0.928	0.927
	0.95	0.972	0.975	0.971	0.914	0.951	0.951	0.959
	0.99	0.985	0.998	0.996	0.919	0.991	0.990	0.989
(100, 50, 50)	0.90	0.943	0.939	0.939	0.889	0.920	0.941	0.924
	0.95	0.970	0.974	0.971	0.906	0.954	0.954	0.960
	0.99	0.984	0.998	0.995	0.912	0.990	0.990	0.990
(100, 100, 50)	0.90	0.904	0.925	0.927	0.914	0.913	0.921	0.934
	0.95	0.950	0.968	0.965	0.951	0.955	0.956	0.968
	0.99	0.991	0.995	0.993	0.976	0.984	0.985	0.993
(100, 100, 100)	0.90	0.910	0.928	0.926	0.915	0.917	0.918	0.931
	0.95	0.955	0.966	0.966	0.958	0.958	0.959	0.967
	0.99	0.990	0.996	0.994	0.979	0.985	0.985	0.992

Mixed distributions. ELP, ELB, IF, PEL, AEL and BTII are competitor approaches.

Table 5. Monte Carlo coverages for the proposed ELQB confidence intervals for θ_{20}, at fixed θ_{10} and θ_{30}.

Sample size	$1-\alpha$	ELQB	ELP	ELB	IF	PEL	AEL	BTII
Scenario 7: $\mathcal{B}(1,6)$, $\mathcal{B}(6,6)$, $\mathcal{B}(9.6,6)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.5$								
(30, 30, 30)	0.90	0.915	0.897	0.926	0.887	0.918	0.934	0.925
	0.95	0.960	0.950	0.969	0.938	0.961	0.969	0.961
	0.99	0.990	0.991	0.994	0.985	0.990	0.993	0.988
(50, 30, 30)	0.90	0.903	0.890	0.918	0.878	0.915	0.932	0.916
	0.95	0.948	0.940	0.958	0.930	0.959	0.968	0.951
	0.99	0.988	0.989	0.994	0.981	0.989	0.994	0.984
(50, 50, 50)	0.90	0.905	0.890	0.913	0.883	0.917	0.928	0.918
	0.95	0.947	0.939	0.954	0.936	0.958	0.966	0.957
	0.99	0.986	0.987	0.992	0.985	0.989	0.992	0.988
(100, 50, 50)	0.90	0.897	0.888	0.906	0.883	0.913	0.923	0.912
	0.95	0.944	0.939	0.954	0.936	0.958	0.966	0.954
	0.99	0.985	0.985	0.991	0.982	0.990	0.992	0.985
(100, 100, 50)	0.90	0.882	0.889	0.903	0.888	0.905	0.913	0.910
	0.95	0.934	0.937	0.949	0.937	0.955	0.961	0.950
	0.99	0.981	0.984	0.986	0.983	0.990	0.992	0.986
(100, 100, 100)	0.90	0.898	0.890	0.908	0.887	0.912	0.920	0.915
	0.95	0.946	0.941	0.951	0.940	0.958	0.962	0.954
	0.99	0.985	0.985	0.989	0.983	0.991	0.992	0.987
Scenario 8: $\mathcal{B}(1,6)$, $\mathcal{B}(9,6)$, $\mathcal{B}(20.4,6)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.8$								
(30, 30, 30)	0.90	0.931	0.904	0.932	0.874	0.913	0.934	0.922
	0.95	0.974	0.961	0.973	0.933	0.956	0.967	0.960
	0.99	0.990	0.994	0.994	0.952	0.991	0.992	0.987
(50, 30, 30)	0.90	0.935	0.907	0.935	0.876	0.917	0.937	0.924
	0.95	0.975	0.963	0.971	0.933	0.958	0.966	0.957
	0.99	0.993	0.995	0.994	0.952	0.990	0.992	0.988
(50, 50, 50)	0.90	0.903	0.893	0.916	0.880	0.912	0.925	0.921
	0.95	0.958	0.946	0.963	0.937	0.955	0.966	0.959
	0.99	0.992	0.993	0.992	0.986	0.992	0.993	0.989
(100, 50, 50)	0.90	0.906	0.894	0.918	0.884	0.914	0.925	0.922
	0.95	0.955	0.948	0.963	0.936	0.957	0.964	0.956
	0.99	0.990	0.989	0.990	0.984	0.990	0.993	0.987
(100, 100, 50)	0.90	0.897	0.898	0.911	0.891	0.909	0.919	0.921
	0.95	0.940	0.945	0.956	0.940	0.958	0.964	0.961
	0.99	0.985	0.988	0.992	0.983	0.991	0.993	0.991
(100, 100, 100)	0.90	0.904	0.893	0.910	0.893	0.906	0.914	0.912
	0.95	0.948	0.944	0.957	0.942	0.955	0.960	0.957
	0.99	0.987	0.990	0.990	0.986	0.990	0.993	0.990
Scenario 9: $\mathcal{B}(1,6)$, $\mathcal{B}(6,6)$, $\mathcal{B}(20.4,6)$, $\theta_{10}=\theta_{30}=0.8$, $\theta_{20}=0.9$								
(30, 30, 30)	0.90	0.912	0.960	0.964	0.830	0.937	0.953	0.952
	0.95	0.925	0.981	0.982	0.851	0.967	0.974	0.977
	0.99	0.935	0.996	0.997	0.874	0.985	0.986	0.993
(50, 30, 30)	0.90	0.910	0.962	0.968	0.815	0.938	0.952	0.944
	0.95	0.924	0.981	0.983	0.831	0.967	0.971	0.968
	0.99	0.933	0.995	0.997	0.858	0.984	0.986	0.989
(50, 50, 50)	0.90	0.954	0.943	0.956	0.906	0.929	0.940	0.942
	0.95	0.974	0.978	0.979	0.923	0.963	0.967	0.969
	0.99	0.986	0.997	0.996	0.934	0.993	0.996	0.994
(100, 50, 50)	0.90	0.949	0.939	0.954	0.889	0.926	0.936	0.935
	0.95	0.972	0.978	0.979	0.905	0.961	0.966	0.968
	0.99	0.983	0.997	0.997	0.916	0.995	0.996	0.992
(100, 100, 50)	0.90	0.921	0.919	0.942	0.907	0.929	0.934	0.942
	0.95	0.966	0.964	0.976	0.946	0.966	0.971	0.974
	0.99	0.994	0.997	0.996	0.974	0.994	0.995	0.995
(100, 100, 100)	0.90	0.916	0.916	0.934	0.914	0.914	0.920	0.933
	0.95	0.957	0.962	0.965	0.949	0.960	0.964	0.966
	0.99	0.993	0.993	0.994	0.981	0.990	0.993	0.991

Beta distributions. ELP, ELB, IF, PEL, AEL and BTII are competitor approaches.

Table 6. Monte Carlo coverages for the proposed ELQB confidence intervals for θ_{20}, at fixed θ_{10} and θ_{30}.

Sample size	$1-\alpha$	ELQB	ELP	ELB	IF	PEL	AEL	BTII
Scenario 10: $0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(2,1)$, $0.5\mathcal{N}(1,1)+0.5\mathcal{N}(4,1.5)$, $0.5\mathcal{N}(3,1.5)+0.5\mathcal{N}(6,1)$								
$\theta_{10}=0.5, \theta_{30}=0.522$, $\theta_{20}=0.674$								
(30, 30, 30)	0.90	0.897	0.851	0.914	0.821	0.911	0.914	0.921
	0.95	0.945	0.921	0.957	0.891	0.932	0.944	0.956
	0.99	0.985	0.975	0.987	0.962	0.976	0.973	0.986
(50, 30, 30)	0.90	0.909	0.859	0.927	0.843	0.899	0.915	0.928
	0.95	0.956	0.928	0.965	0.902	0.919	0.949	0.961
	0.99	0.991	0.981	0.992	0.970	0.975	0.982	0.990
(50, 50, 50)	0.90	0.888	0.840	0.909	0.825	0.864	0.885	0.921
	0.95	0.938	0.907	0.955	0.895	0.902	0.924	0.956
	0.99	0.983	0.973	0.991	0.964	0.956	0.964	0.988
(100, 50, 50)	0.90	0.888	0.842	0.910	0.829	0.847	0.874	0.918
	0.95	0.938	0.907	0.955	0.898	0.897	0.920	0.957
	0.99	0.983	0.973	0.991	0.966	0.956	0.963	0.989
(100, 100, 50)	0.90	0.883	0.830	0.899	0.819	0.862	0.867	0.916
	0.95	0.936	0.898	0.952	0.884	0.898	0.905	0.956
	0.99	0.982	0.975	0.991	0.965	0.946	0.953	0.989
(100, 100, 100)	0.90	0.890	0.834	0.904	0.827	0.851	0.851	0.918
	0.95	0.939	0.905	0.954	0.899	0.891	0.897	0.960
	0.99	0.984	0.977	0.991	0.968	0.946	0.954	0.990

Mixture distributions. ELP, ELB, IF, PEL, AEL and BTII are competitor approaches.

Table 7. Monte Carlo coverages for the proposed ELQB confidence intervals for the VUS.

	$1-\alpha=0.90$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Sample size	ELQB	ELU	JEL	ELQB	ELU	JEL	ELQB	ELU	JEL
Scenario 1: $\mathcal{N}(0,1)$, $\mathcal{N}(2.5,1.1^2)$, $\mathcal{N}(3.69,1.2^2)$, $\gamma_0=0.722$									
(15, 15, 15)	0.887	0.887	0.919	0.935	0.928	0.955	0.982	0.968	0.988
(30, 30, 30)	0.896	0.907	0.910	0.945	0.953	0.954	0.985	0.986	0.992
(50, 30, 30)	0.898	0.906	0.909	0.941	0.953	0.954	0.983	0.986	0.988
(50, 50, 50)	0.899	0.908	0.909	0.946	0.954	0.958	0.987	0.992	0.992
(75, 75, 75)	0.903	0.915	0.915	0.950	0.958	0.960	0.990	0.991	0.993
(100, 100, 100)	0.891	0.903	0.901	0.942	0.949	0.951	0.987	0.988	0.990
Scenario 2: $\mathcal{N}(0,1)$, $\mathcal{N}(3.5,1.1^2)$, $\mathcal{N}(5.5,1.2^2)$, $\gamma_0=0.881$									
(15, 15, 15)	0.889	0.852	0.885	0.946	0.899	0.923	0.979	0.943	0.960
(30, 30, 30)	0.884	0.884	0.895	0.939	0.936	0.946	0.981	0.980	0.980
(50, 30, 30)	0.891	0.887	0.897	0.933	0.937	0.939	0.978	0.979	0.981
(50, 50, 50)	0.888	0.900	0.903	0.943	0.950	0.952	0.983	0.988	0.988
(75, 75, 75)	0.887	0.900	0.902	0.941	0.951	0.951	0.987	0.989	0.990
(100, 100, 100)	0.890	0.904	0.902	0.943	0.954	0.952	0.984	0.991	0.991
Scenario 3: $\mathcal{N}(0,1)$, $\mathcal{N}(4,1.2^2)$, $\mathcal{N}(8.189,2^2)$, $\gamma_0=0.959$									
(15, 15, 15)	0.828	0.731	0.767	0.909	0.785	0.791	0.919	0.829	0.854
(30, 30, 30)	0.836	0.810	0.826	0.898	0.862	0.877	0.960	0.920	0.929
(50, 30, 30)	0.855	0.822	0.840	0.905	0.877	0.887	0.956	0.929	0.935
(50, 50, 50)	0.859	0.851	0.860	0.912	0.901	0.910	0.965	0.959	0.961
(75, 75, 75)	0.874	0.878	0.883	0.922	0.929	0.931	0.973	0.975	0.975
(100, 100, 100)	0.873	0.879	0.881	0.925	0.930	0.936	0.974	0.977	0.979

Normal distributions. ELU and JEL are competitor approaches.

Overall, our approach seems to perform well and is often more accurate than competitors, in all scenarios, particularly when the VUS's true value is large. The ELU method seems to exhibit poor results at low sample sizes.

Table 8. Monte Carlo coverages for the proposed ELQB confidence intervals for the VUS.

	$1-\alpha=0.90$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Sample size	ELQB	ELU	JEL	ELQB	ELU	JEL	ELQB	ELU	JEL
Scenario 4: $\mathcal{G}(6,12)$, $\mathcal{LN}(1.5,0.5)$, $\mathcal{W}(4,6.6)$, $\gamma_0=0.669$									
(15, 15, 15)	0.916	0.933	0.937	0.957	0.967	0.968	0.988	0.987	0.989
(30, 30, 30)	0.897	0.936	0.940	0.950	0.977	0.974	0.990	0.996	0.995
(50, 30, 30)	0.895	0.938	0.938	0.948	0.978	0.975	0.990	0.997	0.995
(50, 50, 50)	0.888	0.932	0.933	0.942	0.974	0.975	0.988	0.996	0.995
(75, 75, 75)	0.870	0.907	0.910	0.928	0.959	0.960	0.983	0.995	0.996
(100, 100, 100)	0.879	0.906	0.910	0.933	0.957	0.961	0.980	0.992	0.993
Scenario 5: $\mathcal{G}(6,12)$, $\mathcal{LN}(1.5,0.5)$, $\mathcal{W}(4,10)$, $\gamma_0=0.868$									
(15, 15, 15)	0.867	0.833	0.861	0.915	0.882	0.906	0.967	0.926	0.947
(30, 30, 30)	0.883	0.883	0.895	0.936	0.936	0.943	0.980	0.980	0.982
(50, 30, 30)	0.878	0.879	0.888	0.933	0.931	0.935	0.978	0.979	0.978
(50, 50, 50)	0.886	0.897	0.899	0.935	0.944	0.948	0.981	0.985	0.987
(75, 75, 75)	0.892	0.901	0.902	0.938	0.950	0.948	0.984	0.989	0.989
(100, 100, 100)	0.888	0.902	0.899	0.937	0.950	0.949	0.984	0.991	0.990
Scenario 6: $\mathcal{G}(6,12)$, $\mathcal{LN}(1.5,0.5)$, $\mathcal{W}(4,12.4)$, $\gamma_0=0.927$									
(15, 15, 15)	0.851	0.765	0.824	0.909	0.821	0.868	0.939	0.873	0.916
(30, 30, 30)	0.861	0.845	0.868	0.912	0.902	0.909	0.966	0.954	0.961
(50, 30, 30)	0.868	0.849	0.864	0.916	0.903	0.913	0.969	0.958	0.961
(50, 50, 50)	0.870	0.872	0.877	0.919	0.923	0.924	0.971	0.970	0.971
(75, 75, 75)	0.883	0.893	0.892	0.933	0.942	0.941	0.980	0.982	0.985
(100, 100, 100)	0.891	0.905	0.906	0.944	0.955	0.953	0.984	0.988	0.989

Mixed distributions. ELU and JEL are competitor approaches.

Table 9. Monte Carlo coverages for the proposed ELQB confidence intervals for the VUS.

	$1-\alpha=0.90$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Sample size	ELQB	ELU	JEL	ELQB	ELU	JEL	ELQB	ELU	JEL
Scenario 7: $\mathcal{B}(1,6)$, $\mathcal{B}(6,6)$, $\mathcal{B}(9.6,6)$, $\gamma_0=0.698$									
(15, 15, 15)	0.867	0.890	0.915	0.923	0.936	0.962	0.984	0.979	0.992
(30, 30, 30)	0.885	0.902	0.906	0.937	0.951	0.954	0.981	0.990	0.993
(50, 30, 30)	0.894	0.914	0.912	0.941	0.956	0.956	0.981	0.991	0.991
(50, 50, 50)	0.896	0.911	0.907	0.945	0.958	0.958	0.987	0.992	0.991
(75, 75, 75)	0.893	0.902	0.902	0.944	0.951	0.951	0.982	0.988	0.988
(100, 100, 100)	0.897	0.910	0.907	0.947	0.955	0.954	0.989	0.991	0.991
Scenario 8: $\mathcal{B}(1,6)$, $\mathcal{B}(9,6)$, $\mathcal{B}(20.4,6)$, $\gamma_0=0.869$									
(15, 15, 15)	0.877	0.840	0.875	0.922	0.888	0.921	0.977	0.937	0.959
(30, 30, 30)	0.891	0.892	0.906	0.941	0.943	0.945	0.983	0.983	0.985
(50, 30, 30)	0.880	0.887	0.889	0.935	0.938	0.943	0.979	0.981	0.982
(50, 50, 50)	0.891	0.902	0.903	0.941	0.955	0.953	0.985	0.988	0.991
(75, 75, 75)	0.891	0.903	0.900	0.943	0.951	0.949	0.984	0.990	0.989
(100, 100, 100)	0.900	0.913	0.911	0.949	0.961	0.957	0.989	0.993	0.993
Scenario 9: $\mathcal{B}(1,6)$, $\mathcal{B}(6,6)$, $\mathcal{B}(20.4,6)$, $\gamma_0=0.917$									
(15, 15, 15)	0.872	0.800	0.849	0.933	0.849	0.892	0.966	0.907	0.941
(30, 30, 30)	0.875	0.856	0.879	0.931	0.913	0.928	0.975	0.962	0.973
(50, 30, 30)	0.889	0.874	0.893	0.939	0.926	0.943	0.985	0.975	0.982
(50, 50, 50)	0.889	0.882	0.897	0.937	0.937	0.943	0.982	0.976	0.983
(75, 75, 75)	0.888	0.892	0.891	0.938	0.939	0.946	0.983	0.982	0.988
(100, 100, 100)	0.882	0.897	0.897	0.939	0.944	0.947	0.985	0.981	0.990

Beta distributions. ELU and JEL are competitor approaches.

Table 10. Monte Carlo coverages for the proposed ELQB confidence intervals for the VUS.

	$1-\alpha=0.90$			$1-\alpha=0.95$			$1-\alpha=0.99$		
Sample size	ELQB	ELU	JEL	ELQB	ELU	JEL	ELQB	ELU	JEL
Scenario 10: $0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(2,1)$, $0.5\mathcal{N}(1,1)+0.5\mathcal{N}(4,1.5)$, $0.5\mathcal{N}(3,1.5)+0.5\mathcal{N}(6,1)$ $\gamma_0=0.544$									
(15, 15, 15)	0.900	0.877	0.921	0.948	0.916	0.965	0.990	0.947	0.994
(30, 30, 30)	0.897	0.898	0.907	0.949	0.941	0.956	0.989	0.978	0.992
(50, 30, 30)	0.896	0.903	0.908	0.944	0.944	0.956	0.988	0.980	0.993
(50, 50, 50)	0.901	0.911	0.911	0.951	0.955	0.960	0.989	0.988	0.993
(75, 75, 75)	0.895	0.904	0.904	0.947	0.951	0.953	0.985	0.985	0.990
(100, 100, 100)	0.896	0.910	0.906	0.947	0.953	0.955	0.988	0.987	0.990

Mixture distributions. ELU and JEL are competitor approaches.

Table 11. Monte Carlo coverages for the proposed ELQB confidence regions for the pair $(\theta_{20}, \theta_{30})$, at fixed θ_{10}.

Scenario	θ_{20}	θ_{30}	(n_1, n_2, n_3)	Nominal level		
				0.90	0.95	0.99
1	0.5	0.8	(20, 20, 20)	0.881	0.945	0.981
			(30, 30, 30)	0.881	0.940	0.990
			(50, 30, 30)	0.877	0.940	0.986
			(50, 50, 50)	0.893	0.947	0.990
			(100, 50, 50)	0.886	0.944	0.991
			(100, 100, 100)	0.883	0.943	0.988
2	0.8	0.8	(20, 20, 20)	0.880	0.943	0.970
			(30, 30, 30)	0.887	0.950	0.989
			(50, 30, 30)	0.865	0.943	0.990
			(50, 50, 50)	0.878	0.939	0.989
			(100, 50, 50)	0.885	0.942	0.988
			(100, 100, 100)	0.883	0.943	0.991
3	0.9	0.9	(20, 20, 20)	0.726	0.744	0.762
			(30, 30, 30)	0.852	0.890	0.909
			(50, 30, 30)	0.859	0.892	0.911
			(50, 50, 50)	0.891	0.947	0.985
			(100, 50, 50)	0.889	0.947	0.984
			(100, 100, 100)	0.879	0.941	0.989

Normal distributions.

3.5 Results: Confidence regions for the pair (TCF_2, TCF_3), at fixed θ_{10}

Finally, Table 11 and Tables S3 to S5 in Section H of the Supplemental Materials report simulation results about confidence regions for the pairs (TCF_2, TCF_3), at fixed $\theta_1 = \theta_{10}$, build using our approach described in Section (2.4). In such simulation experiments, we consider 5000 Monte Carlo replications and some values for the true class fractions in each scenario. Again we set $B = 200$.

As one can see, our method performs well in all considered cases, with the need (as expected) for larger sample sizes (at least (50, 50, 50)) as the true values of the TCFs approach 1.

4 An illustrative example

In the previous sections, we have introduced the theoretical results for tackling the four inferential problems introduced in Section 1. Now, we illustrate how these results can be applied in a practical setting. We aim to show the practical usage of the results we have developed: while the example may not provide a substantial contribution on its own, it serves as a demonstration of how our research can be applied in relevant contexts.

We use a genomic dataset and apply our proposed methods to evaluate the ability of some gene expressions to distinguish DCIS from NC and IBCs. We consider the raw data from series record GSE214540, published on the GEO repository by Guvakova and Sokol.[22] The raw data contains mRNA expression levels of different genes from formalin-fixed paraffin-embedded (FFPE) human breast tissue samples generated by using the QuantiGene Plex 2.0 assay and Flex-Map 3D.[23] The

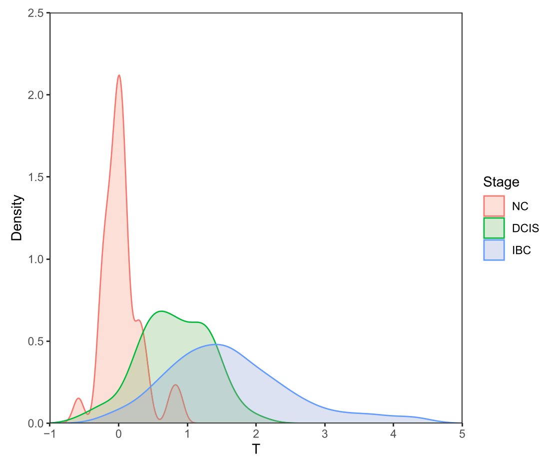

Figure 1. Estimated densities for the combination $T = \text{KRT8} - 0.9 \times \text{KRT5} + 0.3 \times \text{CDH2}$.

FFPE tissues were collected from the Department of Pathology and Laboratory Medicine, Tumor Tissue and Biospecimen Bank, and the Cooperative Human Tissue Network at the University of Pennsylvania.

Guvakova and Sokol[22] measured 14 target genes, namely: estrogen receptor 1 (ESR1), progesterone receptor (PGR), erb-b2 receptor tyrosine kinase 2 (ERBB2), insulin-like growth factor 1 receptor (IGF1R), vav guanine nucleotide exchange factor 1 (VAV1), vav guanine nucleotide exchange factor 2 (vav guanine nucleotide exchange factor 2), vav guanine nucleotide exchange factor 3 (VAV3), ras-related protein Rap-1A (RAP1A), ras-related protein Rap-1b (RAP1B), Rap guanine nucleotide exchange factor 1 (RAPGEF1), keratin 5 (KRT5), keratin 8 (KRT8), cadherin 1 (CDH1), cadherin 2 (CDH2); and two housekeeping genes, namely, peptidylprolyl isomerase B (PPIB) and glucuronidase beta(GUSB). As noted by Prabakaran et al.,[23] in FFPE tissue, PPIB expressed consistently, whereas GUSB showed a relatively low expression level. For this reason, we do not take into account the mRNA expression levels of GUSB in our analysis.

Before doing the analysis, the normalization of the raw mRNA data is required. Within our analysis, the normalized data are obtained through three steps: firstly, we subtract background values from each measurement, to obtain the real values of gene expressions; secondly, we obtain the housekeeping normalization factor by dividing the average of housekeeping gene values, the PPIB, to each of its values; then, the normalized mRNA values of each gene is obtained as a ratio of mRNA values and the housekeeping normalization factor. Note that, in the first step, negative values are set as 0, to reflect the fact that there is no measurable expression of the target gene in the assay. We then apply $\log_{10}(x + 1)$ transformation to the normalized values for the main analysis.

The final dataset contains the $\log_{10}$-transformed (normalized) mRNA expression values of 14 target genes from FFPE tissues of 251 women: 34 for the NC group, 75 for the DCIS group, and 142 for the IBC group. A preliminary analysis indicates that almost all genes have poor accuracy in classifying the three stages of breast cancer (see VUS estimates in Table S6, Section I of the Supplemental Materials). Hence, we consider a linear combinations, say T, of the $\log_{10}$-transformed mRNA expression levels for three genes: KRT5, KRT8, and CDH2, i.e. $T = \text{KRT8} - 0.9 \times \text{KRT5} + 0.3 \times \text{CDH2}$. The coefficients of such a combination are based on the maximization process of the estimated VUS, as in Zhang and Li.[24] Then, in the analysis, we treat this combination as exogenously fixed.

The estimated VUS for the combined test T is 0.685. Employing our method ELBQ, proposed in Section 2.3, the 90%, 95%, and 99% EL confidence intervals for the VUS are $(0.626, 0.740)$, $(0.614, 0.750)$ and $(0.591, 0.769)$, respectively. By these results, T seems to have a sufficiently good capacity for discrimination among the three stages of breast cancer.

For the three stages, the kernel-based estimated densities of T are shown in Figure 1. By inspection of the figure, we choose two plausible values for the thresholds t_1 and t_2: 0.275 and 1.35, respectively. These values roughly correspond to the crossing points of the estimated densities of the test T in the groups, a choice that guarantees some optimality, at least from an empirical point of view. In practice, to select such thresholds, researchers may conduct preliminary experiments or review existing literature to identify values that are commonly used or have demonstrated good performance in similar studies. This would ensure that the selected thresholds are reasonable and align with established practices in the field. Treating values of t_1 and t_2 as fixed, Figure 2 shows the corresponding 95% confidence region for the TCFs $(\theta_{10}, \theta_{20}, \theta_{30})$ on the estimated ROC surface, obtained by our ELQ3D method in Section 2.1. Moreover, if we fix $\theta_{10} = 0.8$ and $\theta_{30} = 0.6$, the empirical estimate $\widehat{\theta}_2$ is about 0.707, and the 95% ELQB (Section 2.2, with 200 bootstrap replications) confidence interval for the probability of correct classification in the DCIS stage is $(0.461, 0.889)$. The width of such an interval reflects the

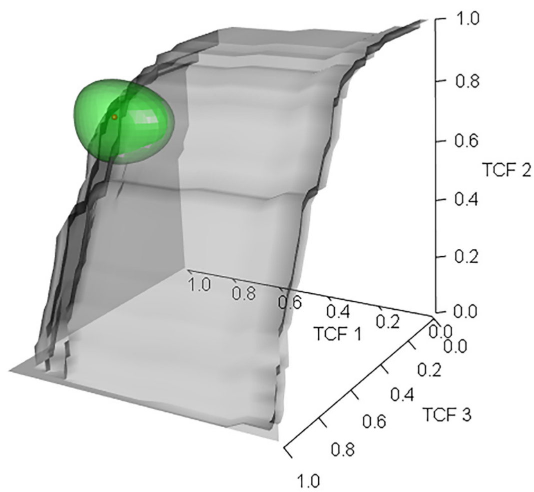

Figure 2. Empirical receiver operating characteristic (ROC) surface for T, and the confidence region for $(\theta_{10}, \theta_{20}, \theta_{30})$ when the pair (t_1, t_2) is (0.275, 1.35). The point estimate $(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3)$ is (0.824, 0.747, 0.578).

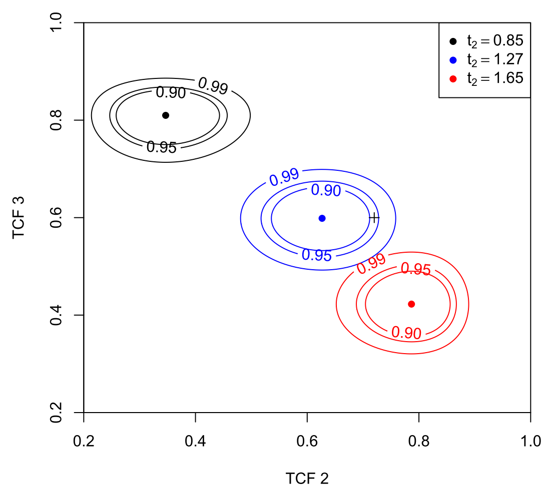

Figure 3. Confidence regions for $(\theta_{20}, \theta_{30})$ when $\theta_{10} = 0.9$, for three different choices of t_2: 0.85, 1.27 and 1.65. The point estimates $(\widehat{\theta}_2, \widehat{\theta}_3)$ are (0.347, 0.810), (0.627, 0.599) and (0.787, 0.423), respectively. Symbol "+" is at $\mathrm{TCF}_2 = 0.72$ and $\mathrm{TCF}_3 = 0.6$.

variability of the data and the degree of overlap among the estimated densities of the combination T, in particular between the DCIS and IBC stages. Finally, Figure 3 shows ELQB confidence regions (Section 2.4, with 200 bootstrap replications) for the pair $(\theta_{20}, \theta_{30})$ when $\theta_{10} = 0.9$ and the threshold t_2 is chosen to be 0.85, 1.27, and 1.65, respectively. Confidence regions indicate, at different levels, the pairs (θ_2, θ_3) which are compatible with the constraint $\theta_{10} = 0.9$ for the combination T, for three possible choices of threshold t_2. Then, for example, Figure 3 indicates that, at level 0.95, the combination T may perform with values for TCF_2 and TCF_3 equal to 0.72 and 0.6, respectively when we require $\mathrm{TCF}_1 = 0.9$ and use $t_2 = 1.27$. These regions could therefore serve as valuable indicators for the selection process related to t_2 and could provide valuable insights for decision-making.

5 Discussion

We present a fairly general approach for constructing confidence intervals and regions in a three-class ROC analysis. Our approach allows getting adequate techniques to solve inferential problems concerning the evaluation of a diagnostic test (or a biomarker) and, in particular, to obtain: (a) confidence regions for the triplet $(\mathrm{TCF}_1, \mathrm{TCF}_2, \mathrm{TCF}_3)$ corresponding to a specific choice for the thresholds t_1 and t_2; (b) confidence intervals for the VUS; (c) confidence intervals for the probability of correct classification to the "early stage," TCF_2, for fixed TCF_1 and TCF_3; (d) confidence regions for a pair of TCFs, when it is fixed the value of the remaining third. The proposed techniques are justified theoretically and are validated empirically using a large simulation study, where they are also compared with competitors, if present in the literature. Overall, simulation results reveal that the proposed methods perform well in general, and are at least as accurate as competitors, showing better performance in several situations. The R codes for implementing the proposed approaches are available online.

It is noteworthy that while only one of the four above-mentioned inferential problems could be of immediate interest in practice, presenting a unifying approach that encompasses all the situations allows a more holistic view of the problem of evaluation of a diagnostic test and makes various extensions amenable, as briefly discussed below.

Confidence intervals for the hypervolume under the ROC manifold. Suppose that the disease is articulated according to $M > 3$ ordered stages. Let $Y_j, j = 1, \dots, M$, be the test results for a subject within jth class. In such a situation, the hypervolume under the ROC manifold (HUM), extends the concept of VUS and is defined as $\beta = \Pr(Y_1 < Y_2 < \dots < Y_M)$. An estimator, consistent and asymptotically normal,[7] of β can be obtained as:

$$\hat{\beta} = \widehat{\Pr}(Y_1 < Y_2 < \dots < Y_M) = \frac{1}{n_1 n_2 \dots n_M} \sum_{i=1}^{n_1} \sum_{r=1}^{n_2} \dots \sum_{k=1}^{n_M} \mathrm{I}\left(Y_{1i} < Y_{2r} < \dots < Y_{Mk}\right).$$

Since β is a probability, we can write the EL statistic $\ell(\beta)$ as in (9), and by generalizing Theorem 2.3, can prove that $\ell(\beta_0)$ has asymptotically a scaled χ_1^2 distribution, under some weak conditions. Thus, a confidence interval for the HUM can be obtained by

$$\mathcal{R}_{\beta,\alpha} = \left\{ \beta : \hat{w}\ell(\beta) \leq \chi^2_{1,(1-\alpha)} \right\}.$$

Confidence intervals for the covariate-specific VUS. Often, in biomedical studies, the researchers collect not only results of potential diagnostic tests but also additional information, about the subjects under study, as covariates (e.g. age, sex, comorbidity profiles). In such cases, covariate-specific measures to evaluate the accuracy of the diagnostic tests are relevant. For the three-class setting, covariate-specific VUS estimators are proposed in To et al.[25] Such estimators are consistent and asymptotically normal. If, for a given vector of covariate values x, $\hat{\gamma}(x)$ denotes an estimate, a confidence interval for the covariate-specific VUS, $\gamma(x)$, can be obtained again by (9), where γ and $\hat{\gamma}$ are replaced by $\gamma(x)$ and $\hat{\gamma}(x)$, respectively.

Confidence regions for optimal thresholds and associated TCF. Although scarcely used, a criterion for choosing an optimal threshold in a two-class setting, is the so-called symmetric point (see Lòpez-Ratón[26] and references therein). This approach finds the threshold t at which the sensitivity and specificity of the diagnostic test have the same value. In the three-class setting, the extension of the symmetric point approach is trivial: the optimal thresholds t_1, t_2 are such that TCF_1, TCF_2, and TCF_3 have the same value, i.e. $\theta_1 = \theta_2 = \theta_3$. Because this criterion imposes two constraints on θ_1, θ_2 and θ_3, i.e. $\theta_1 = \theta_2$ and $\theta_2 = \theta_3$, if we denote by θ the common value and, from (5), let $\ell_+(\theta, t_1, t_2) = \ell(\theta, \theta, \theta; t_1, t_2)$, we can prove that $\ell_+(\theta, t_1, t_2)$ again has an asymptotic χ_3^2 distribution, under the true parameters values. Therefore, ℓ_+ can be used to build confidence regions for the symmetric point-based optimal thresholds and the associated common value of TCFs. This technique extends the proposal discussed in Adimari and Sinigaglia[27] for two classes, to the three-class setting.

An interesting topic that remains to be developed concerns the problem of building confidence regions for optimal TCFs, i.e. TCFs corresponding to thresholds chosen through other criteria, such as the one based on the generalized Youden index,[28] the closest to perfection and the max volume.[29] Such a topic will be the focus of future work.

Acknowledgments

The insightful feedback and suggestions of three reviewers in enhancing the quality of the paper are gratefully acknowledged.

Data availability

Software in the form of R codes is available on https://github.com/toduckhanh/emplikROCS.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author was supported by the Ministero dell'Istruzione, dell'Universitàe della Ricerca-Italy (grant number DIFO_ECCELLENZA18_01), while being an assistant professor at the Department of Statistical Sciences of the University of Padova, Italy.

ORCID iDs

Duc-Khanh To https://orcid.org/0000-0002-4641-0764
Gianfranco Adimari https://orcid.org/0000-0002-7811-912X

Supplemental material

Supplemental material for this article is available online.

References

1. Owen AB. *Empirical likelihood*. New York: Chapman and Hall/CRC, 2001.
2. Hjort NL, McKeague IW and van Keilegom I. Extending the scope of empirical likelihood. *Ann Stat* 2009; **37**: 1079–1111.
3. Adimari G and Guolo A. A note on the asymptotic behaviour of empirical likelihood statistics. *Stat Method Appl* 2010; **19**: 463–476.
4. Lazar NA. A review of empirical likelihood. *Annu Rev Stat Appl* 2021; **8**: 329–344.
5. Liu P and Zhao Y. A review of recent advances in empirical likelihood. *WIREs: Comput Stat* 2023; **15**: e1599.
6. Nakas CT. Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *REVSTAT - Stat J* 2014; **12**: 43–65.
7. Nakas CT and Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Stat Med* 2004; **23**: 3437–3449.
8. Xiong C, van Belle G, Miller JP et al. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Stat Med* 2006; **25**: 1251–1273.
9. Li J and Zhou XH. Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *J Stat Plan Infer* 2009; **139**: 4133–4142.
10. Kang L and Tian L. Estimation of the volume under the ROC surface with three ordinal diagnostic categories. *Comput Stat Data An* 2013; **62**: 39–51.
11. Guangming P, Xiping W and Wang Z. Nonparametric statistical inference for $P(X < Y < Z)$. *Sankhya A* 2013; **75**: 118–138.
12. Jing BY, Yuan J and Zhou W. Jackknife empirical likelihood. *J Am Stat Assoc* 2009; **104**: 1224–1232.
13. Wan S. An empirical likelihood confidence interval for the volume under ROC surface. *Stat Probabil Lett* 2012; **82**: 1463–1467.
14. Dong T, Tian L, Hutson A et al. 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. *Stat Med* 2011; **30**: 3532–3545.
15. Dong T and Tian L. Confidence interval estimation for sensitivity to the early diseased stage based on empirical likelihood. *J Biopharm Stat* 2015; **25**: 1215–1233.
16. Rahman H, Zhao Y and Initiative ADN. Empirical likelihood confidence interval for sensitivity to the early disease stage. *Pharm Stat* 2022; **21**: 566–583.
17. Hai Y, Shi S and Qin G. Bayesian and influence function-based empirical likelihoods for inference of sensitivity to the early diseased stage in diagnostic tests. *Biometrical J* 2023; **65**: 2200021.
18. Adimari G. An empirical likelihood statistic for quantiles. *J Stat Comput Sim* 1998; **60**: 85–95.
19. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.
20. Feng D and Tierney L. Computing and displaying isosurfaces in R. *J Stat Softw* 2008; **28**: 1–14.
21. Lehmann EL. *Elements of Large-Sample Theory*. New York: Springer Science & Business Media, 1998.
22. Guvakova MA and Sokol S. The g3mclass is a practical software for multiclass classification on biomarkers. *Sci Rep-UK* 2022; **12**: 18742.
23. Prabakaran I, Wu Z, Lee C et al. Gaussian mixture models for probabilistic classification of breast cancergaussian mixture models for cancer classification. *Cancer Res* 2019; **79**: 3492–3502.
24. Zhang Y and Li J. Combining multiple markers for multi-category classification: an ROC surface approach. *Aust NZ J Stat* 2011; **53**: 63–78.
25. To DK, Adimari G and Chiogna M. Estimation of the volume under a ROC surface in presence of covariates. *Comput Stat Data An* 2022; **174**: 107434.
26. López-Ratón M, Cadarso-Suárez C, Molanes-López EM et al. Confidence intervals for the symmetry point: an optimal cutpoint in continuous diagnostic tests. *Pharm Stat* 2016; **15**: 178–192.
27. Adimari G and Sinigaglia A. Nonparametric confidence regions for the symmetry point-based optimal cutpoint and associated sensitivity of a continuous-scale diagnostic test. *Biometrical J* 2020; **62**: 1463–1475.
28. Nakas CT, Alonzo TA and Yiannoutsos CT. Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Stat Med* 2010; **29**: 2946–2955.
29. Attwood K, Tian L and Xiong C. Diagnostic thresholds with three ordinal groups. *J Biopharm Stat* 2014; **24**: 608–633.