Point and interval estimation of accuracies of a binary medical diagnostic test following group sequential testing

BY YU SHU1,*, AIYI LIU2 AND ZHAOHAI LI2,3

1 Global Pharmaceutical Research and Development, Statistics, Abbott Laboratories, Building AP9A-1, 100 Abbott Park Road, Abbott Park, IL 60064, USA
2 Biometry and Mathematical Statistics Branch, National Institute of Child Health and Human Development, 6100 Executive Boulevard, Rockville, MD 20852, USA
3 Department of Statistics, George Washington University, 2140 Pennsylvania Avenue Northwest, Washington, DC 20037, USA

When hypotheses concerning the sensitivity and specificity of a binary medical diagnostic test are simultaneously tested using a group sequential procedure, constructing point and interval estimates of the parameters is challenging because there is no unique way to order sample points in the two-dimensional space. In this paper, upon termination of a group sequential procedure, we compare the bias and mean squared errors of the maximum-likelihood and Rao–Blackwell unbiased estimators of sensitivity and specificity. Confidence intervals (CIs) of the two parameters were constructed using normal approximation and Woodroofe’s pivot methods based on maximum-likelihood and Rao–Blackwell unbiased estimates. The coverage probability and the expected length of CIs for the parameters were compared by simulation studies.

Keywords: group sequential tests; sensitivity; specificity; Woodroofe pivot method

1. Introduction

Accurate diagnosis of a medical condition is often the first step towards cure of the disease. Sensitivity (se) and specificity (sp), the probability of correctly classifying a true case of disease as diseased and a non-case as non-diseased, respectively, are two important measures to evaluate a medical diagnostic test (Zhou et al. 2002). Although medical test results are frequently measured on a continuous scale or in ordinal categories, they can always be dichotomized to binary outcomes by choosing a threshold value, on which we focus here. In assessing the performance of a diagnostic test, we want to know whether the test results differentiate the two health states, presence and absence of the medical condition.

* Author for correspondence (yu.shu@abbott.com).

One contribution of 13 to a Theme Issue ‘Mathematical and statistical methods for diagnoses and therapies’.
Group sequential testing procedures have been considered to evaluate a diagnostic test and to estimate parameters of accuracy. Obuchowski & Zhou (2002) considered two-stage designs in which all subjects tested positive in the first stage; a random sample of test negatives are selected for disease verification in the second stage to reduce the costs of prospective studies of binary diagnostic tests. Wruck et al. (2006) proposed estimators based on a two-step sequential sampling scheme in a large cohort study, in which the first step of sampling is conducted to efficiently estimate specificity, and the second step is to estimate sensitivity if the candidate test is determined to have sufficiently high specificity. The sensitivity and the specificity have also been evaluated simultaneously in a group sequential design with a case-control sampling method under the criteria of minimizing the expected sample size when the candidate diagnostic test performs poorly (Shu et al. 2007).

In medical studies, a more complete analysis is usually required than simply the ‘accept’ or ‘reject’ decision of a hypothesis. Medical investigators are interested in how accurately a medical test can distinguish diseased from disease-free subjects in the general population. An interval estimate for the measures of accuracy will allow consideration of the magnitude of the accuracy and indicate its practical importance, irrespective of whether a hypothesis test declares statistical significance.

For sequential sampling with Bernoulli trials, Girshick et al. (1946) and Lehmann & Stein (1950) considered unbiased estimation of the success probability and further found necessary and sufficient conditions for completeness, enabling conclusions about uniformly minimum variance unbiased estimators. Liu et al. (2006) considered unbiased estimation following a group sequential test for distributions in a one-parameter exponential family. Jennison & Turnbull (1983) described a method for constructing CIs for a binomial parameter following a multistage test. The method involves defining an ordering among points on the stopping boundaries and computing the probability of a point being more extreme in this ordering than the observed one. Whether or not the ordering of the sample space may be reasonable depends on the particular stopping rule used in the sequential testing. Todd & Whitehead (1997) extended Woodroffe’s (1992) pivot technique to provide CIs for the absolute success probability of two treatments in a clinical trial.

In this paper, we consider the joint inference of sensitivity and specificity following a group sequential test. In §2, we derive the distribution theory of the empirical sensitivity and specificity upon termination of the group sequential testing. In §3, the maximum-likelihood and Rao–Blackwell unbiased estimates are obtained and compared in terms of their bias and mean squared error. Adopting Woodroffe’s pivot method, approximate CIs of sensitivity and specificity are constructed in §4. Simulation results are presented to compare the performance of different CIs. Some discussions are given in §5.

2. Distribution theory

Suppose that the hypotheses concerning the sensitivity and specificity of a binary diagnostic test are

\[ H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1, \]

\[ (2.1) \]
where \( \theta = (\text{se}, \text{sp})^T \) is a column vector (the superscript ‘T’ denotes the transpose of a vector), and \( \Theta_i, i=0, 1, \) are the parameter spaces under \( H_i, i=0, 1. \) The multistage testing procedure with \( K \) interim analyses is as follows. At the \( k \)th stage, \( 1\leq k<K, \) if \( S_k \in B_k, \) stop the sampling and reject \( H_0; \) otherwise, if \( S_k \in C_k, \) continue the sampling to the \((k+1)\)th stage. Here \( S_k \) is a two-dimensional vector, with the two elements representing, respectively, the cumulative number of correctly classified diseased and non-diseased subjects up to stage \( k, \) and \( B_k \) and \( C_k \) are the stopping and continuation regions at stage \( k, \) respectively.

Upon termination of the sequential testing, we observe two statistics: \( M, \) the (random) number of analyses performed, and \( S_M. \) Since, for each fixed \( n, S_n \) is sufficient for \( \theta, \) the statistics \( (M, S_M) \) is jointly sufficient for \( \theta \) with respect to the sample space

\[
\{ S_1 \in B_1 \} \cup \bigcup_{k=2}^{K} \{ S_i \in C_i, i<k, S_k \in B_k \}.
\]

Thus, estimation of \( \theta \) can be solely based on \( (M, S_M). \)

Denote by ‘+’ and ‘−’ the diseased and non-diseased groups, respectively, by \( n_k = (n_k^+, n_k^-)^T \) the vector representing the cumulative sample sizes (for diseased and non-diseased) up to stage \( k, \) and by \( m_k = (m_k^+, m_k^-)^T = n_k - n_{k-1} \) (\( n_0 = 0 \)) the vector representing the incremental sample sizes (for diseased and non-diseased) at stage \( k, \) respectively. To obtain the joint distribution function of \( (M, S_M), \) we first note that the increments \( X_1 = S_1, \) \( X_2 = S_2 - S_1, \ldots, X_k = S_k - S_{k-1} \) are independently distributed with joint density

\[
P_\theta \{ X_1 = (X_1^+, X_1^-)^T, \ldots, X_k = (X_k^+, X_k^-)^T \} = \prod_{i=1}^{k} b(X_i^+, m_i^+; \text{se})b(X_i^-, m_i^-; \text{sp}),
\]

(2.2)

where \( b(a,m,p) \) is the binomial density function with \( a \) successes in \( m \) independent trials with success probability \( p. \) Using this density function we can easily conclude that the density function of the sufficient statistics \( (M, S_M) \) at \( (k, s) \) is

\[
g_\theta(k, s) = r_k(s) f_k; \theta(s),
\]

(2.3)

where \( f_k; \theta(s) = b(S^+, n_k^+, \text{se})b(S^-, n_k^-, \text{sp}), \) \( r_1(s) = 1_{B_1}(s) \) and, for \( k>1, \)

\[
r_k(s) = \sum_{u_1} \cdots \sum_{u_{k-1} \in C_i, i<k} \tilde{r}_k(s; u_1, u_2, \ldots, u_{k-1}),
\]

(2.4)

with

\[
\tilde{r}_k(s_k; s_1, \ldots, s_{k-1}) = \frac{1}{h_{s_k}(s_k)} h_{m_1}(s_1)h_{m_2}(s_2 - s_1)\ldots h_{m_k}(s_k - s_{k-1}),
\]

where for \( n = (n^+, n^-)^T \) and \( s = (S^+, S^-)^T \)

\[
h_n(s) = C_{n^+}^{S^+} \times C_{n^-}^{S^-},
\]

Here, \( C_{m}^{n} \) is the combinatorial number of choosing \( m \) from \( n. \)
Formula (2.3) provides the basis for computing the expectation and variance of a statistic \( W = w(M, S_M) \):

\[
E_{\theta}(W) = \sum_{k=1}^{K} \sum_{s} w(k, s) r_k(s) f_{k, \theta}(s), \tag{2.5}
\]

\[
\text{Var}_{\theta}(W) = \sum_{k=1}^{K} \sum_{s} [w(k, s) - E_{\theta}(W)]^2 r_k(s) f_{k, \theta}(s). \tag{2.6}
\]

### 3. Point estimation

(a) Maximum-likelihood estimates

Suppose that a null hypothesis concerning the value of a parameter \( \theta \) is tested using a sequential procedure; the maximum-likelihood estimator \( \hat{\theta} \) of \( \theta \) is calculated from the data collected. Its value will not be altered by the fact that the test was conducted sequentially; however, the distribution of \( \hat{\theta} \) will be affected in a substantial way. Estimators that are unbiased or almost unbiased in fixed size sampling usually yield substantial bias if used with sequential sampling.

The likelihood function of \( \theta \) is

\[
L(\theta; M, S_M) = r_M(s)f_{M, \theta}(s). \tag{3.1}
\]

In equation (3.1), only \( f_{M, \theta}(s) \) involves the parameters. Maximizing the likelihood function, we have

\[
\hat{\theta} = \left( \begin{array}{c}
\hat{\theta}_1 \\
\hat{\theta}_2
\end{array} \right) = \left( \begin{array}{c}
\hat{\text{se}} \\
\hat{\text{sp}}
\end{array} \right) = \left( \begin{array}{c}
S_{M}^+ / n_M^+ \\
S_{M}^- / n_M^-
\end{array} \right). \tag{3.2}
\]

Thus, the values of the maximum-likelihood estimates of sensitivity and specificity will not be altered by the test being conducted sequentially.

However, these estimators are no longer unbiased. The expectation of \( \hat{\theta} \) in two-stage designs can be expressed as follows:

\[
E_{\theta}[\hat{\text{se}}] = \sum_{s_1 \in B_1} \frac{S_{1}^+}{n_1^+} f_{1, \theta}(s_1) + \sum_{s_2 \in C_1} \sum_{u \in \tilde{C}_1} \frac{S_{2}^+}{n_2^+} \tilde{r}_2(s)f_{2, \theta}(s_2), \tag{3.3}
\]

\[
E_{\theta}[\hat{\text{sp}}] = \sum_{s_1 \in B_1} \frac{S_{1}^-}{n_1^+} f_{1, \theta}(s_1) + \sum_{s_2 \in C_1} \sum_{u \in \tilde{C}_1} \frac{S_{2}^+}{n_2^-} \tilde{r}_2(s)f_{2, \theta}(s_2). \tag{3.4}
\]

Further simplification of the above equations depends on the stopping and continuation regions at each stage. Unlike in fixed size sampling, it is apparent that the biases of the maximum-likelihood estimates are functions of both sensitivity and specificity, which is understandable since the stopping of the sampling depends on both parameters.
The variance expressions of $\hat{\theta}_i$, $i=1, 2$, for a two-stage design are given by
\[
\text{Var}(\hat{\theta}_i) = \sum_{s_1 \in B_1} [\hat{\theta}_i - E(\hat{\theta}_i)]^2 f_{1, \theta}(s_1) + \sum_{s_2} \sum_{u \in C_1} [\hat{\theta}_i - E(\hat{\theta}_i)]^2 \hat{r}_2(s_2)f_{2, \theta}(s_2).
\]

(b) Rao–Blackwell unbiased estimates

Let $\hat{\eta}_1(S_1)$ be the unique sufficient statistic-based unbiased estimator, which coincides with the maximum-likelihood estimator, of $\eta$ for a non-sequential sampling with size $n_1 = (n_1^+, n_1^-)^T$. Then the Rao–Blackwell estimator is given by
\[
\hat{\eta}_u(M, S_M) = E[\hat{\eta}_1(S_1)|(M, S_M)],
\]
which is unbiased for $\eta$ and with reduced variance. A recursive expression for the Rao–Blackwell estimator is $\hat{\eta}_u(1, S_1) = S_1/n_1$ and, for $k>1$,
\[
\hat{\eta}_u(k, S_k) = \frac{1_{B_k}(S_k)}{r_k(S_k)h_{n_k}(S_k)} \sum_{u \in C_{k-1}} \hat{\eta}_u(k-1, u)r_{k-1}(u)h_{n_{k-1}}(u)h_{m_k}(S_k - u).
\]

In a two-stage design, the Rao–Blackwell unbiased estimator at the second stage is a weighted average of the first-stage estimator,
\[
\hat{\eta}_u(2, S_2) = \frac{\sum_{u \in C_1} \hat{\eta}_u(1, u)h_{n_1}(u)h_{m_1}(S_2 - u)}{\sum_{u \in C_1} h_{n_1}(u)h_{m_1}(S_2 - u)}.
\]

The variance of the Rao–Blackwell estimator can be computed using (2.6).

(c) Comparison of estimators

Shu et al. (2007) proposed an optimal two-stage test to evaluate sensitivity and specificity of a binary medical test simultaneously. The objective was to minimize the expected sample size when the accuracies of a medical test are below a desired level of expectation. Let $a_1^+$ and $a_1^-$ be the stopping boundaries for test statistics $S_1^+$ and $S_1^-$ in the first stage. If the test statistics exceed the stopping boundaries, then the process continues to the second stage; otherwise it stops for low sensitivity or specificity.

Suppose that $H_0 : \theta_0 = (s_{0,0}, s_{0,1})^T = (0.7, 0.75)^T$. Then the power of the test is evaluated at $\theta_1 = (0.9, 0.95)^T$ and the probabilities of type I and II errors are $\alpha = 0.05$, $\beta = 0.1$, respectively. The optimal two-stage design requires 14 diseased and 12 non-diseased subjects in stage 1, and a total of 32 subjects per group at the end of stage 2. Only when both the test statistics $(S_1^+, S_1^-)$ exceed $(10, 9)$ in stage 1, and $(S_2^+, S_2^-)$ exceed $(25, 27)$ in stage 2, is the null hypothesis rejected (see Shu et al. 2007). It is clear that the stopping and continuation regions depend on both test statistics. Consequently, the bias, variance and mean squared error of the estimates are functions of sensitivity and specificity. Figure 1 shows the bias of the maximum-likelihood estimate in this optimal two-stage design.

Firstly, we observe that the negative biases shown in figure 1 indicate that the maximum-likelihood estimators underestimate both sensitivity and specificity. This negativity of the bias is attributed to the design features, where only one-sided lower stopping boundaries are being used, and the sequential testing only stops early under the null hypothesis. (Liu et al. (2005) noted similar patterns in the bias of the maximum-likelihood estimators for certain optimal two-stage
designs for phase II clinical trials.) Secondly, figure 1 reveals a consistent trend that the magnitude of the bias of the maximum-likelihood estimator of one parameter (sensitivity or specificity) increases as the value of the other parameter increases. For example, in figure 1a, the maximum absolute bias of the maximum-likelihood estimator of sensitivity is close to 0.025 when the true value of specificity is 0.9, and decreases to 0.015 when specificity is 0.8; the bias is negligible when specificity is less than 0.5. Thirdly, the absolute bias appears to reach its maximum at around the values of the parameters under the null hypothesis (se0 = 0.7 and sp0 = 0.75).

Figure 2 shows the square root of the mean squared error of the maximum-likelihood and the Rao–Blackwell unbiased estimators of sensitivity and specificity in the optimal two-stage design. The difference in mean squared error between the maximum-likelihood and the Rao–Blackwell estimators of the parameters is more obvious when the parameters of interest have relatively large values. The value of the other parameter has more effect on the mean squared error of the maximum-likelihood estimator of one parameter than on that of the Rao–Blackwell estimator of the same parameter. Although the curves in figure 2a,b show that the mean squared error of the maximum-likelihood estimator of sensitivity decreases more than that of the Rao–Blackwell estimator when specificity increases, the magnitude of the reduction is not appealing taking into account the bias of the maximum-likelihood estimator. The same pattern is observed when comparing figure 2c,d for specificity.

4. CIs for sensitivity and specificity

(a) CIs based on pivotal quantities

Woodroffe (1992) proposed a simple approach for the problem of estimating a normal mean θ following a truncated sequential probability ratio test and provided a useful method for constructing confidence bounds and intervals based
on an approximate pivot. The CIs of sensitivity and specificity can also be obtained using this approximate pivot method. The method is based on constructing an approximately normally distributed pivotal quantity and then employing traditional probability arguments.

Continuing to use the notation of the previous sections, $W_1(\text{se})$ and $W_2(\text{sp})$ are defined as

$$W_1(\text{se}) = \frac{S_M^+ - \bar{n}_M^+\text{se}}{\sqrt{n_M^+}},$$

$$W_2(\text{sp}) = \frac{S_M^- - \bar{n}_M^-\text{sp}}{\sqrt{n_M^-}}.$$

In the group sequential designs, $n_M^+$ and $\bar{n}_M$ are random variables depending on the test statistics in previous stages, $S_k^+$ and $S_k^-$, $k < M$. Hence, $W_1(\text{se})$ and $W_2(\text{sp})$ are no longer normally distributed with zero mean and unit variance. Without loss of generality, we consider the CI of sensitivity. Let $\delta(\text{se})$ and $\sigma(\text{se})$ be the mean and standard deviation functions of $W_1(\text{se})$, respectively. Then $W'_1(\text{se})$ is defined as

$$W'_1(\text{se}) = \frac{W_1(\text{se}) - \delta(\text{se})}{\sigma(\text{se})}. \quad (4.1)$$

Figure 2. Square root of mean squared error of (a,c) the maximum likelihood and (b,d) Rao–Blackwell unbiased estimators of both (a,b) sensitivity and (c,d) specificity in an optimal two-stage design. (a,b) Specificity: solid curves, 0.1; long dashed curves, 0.6; dot-dashed curves, 0.7; dotted curves, 0.8; short dashed curves, 0.9. (c,d) Sensitivity: solid curves, 0.1; long dashed curves, 0.6; dot-dashed curves, 0.7; dotted curves, 0.8; short dashed curves, 0.9.
When the sample size is sufficiently large, this quantity has zero mean and unit standard deviation and is treated by approximately following the standard normal distribution. Then an approximate \((1 - \alpha)\)-level CI for sensitivity is of the form

\[
\left\{ \frac{S^+_M}{n^+_M} - \frac{\hat{\delta}(se)}{\sqrt{n^+_M}} + \frac{\sigma(se)}{\sqrt{n^+_M}} Z_{\alpha/2} \right\},
\]

(4.2)

where \(\hat{\delta}(se)\) and \(\hat{\sigma}(se)\) are proper estimates of \(\delta(se)\) and \(\sigma(se)\), respectively. Note that

\[
\hat{\delta}(se) = EW_1(se) = E \left[ \frac{S^+_M - n^+_M se}{\sqrt{n^+_M}} \right] = \sum_{k=1}^{K} s^+ - n^+_M se \ g_\theta(k, s),
\]

where \(s = (s^+, s^-)\). Then an estimate of \(\hat{\delta}(se)\) can be given by

\[
\hat{\delta}(se) = \sum_{k=1}^{K} \sum_{s \in B_k} s^+ - n^+_M se \ g_\theta(k, s).
\]

Similarly \(\hat{\sigma}(se)\) can be estimated by

\[
\hat{\sigma}(se) = \sum_{k=1}^{K} \sum_{s \in B_k} \left( \frac{s^+ - n^+_M se}{\sqrt{n^+_M}} \right)^2 g_\theta(k, s) - \left[ \hat{\delta}(se) \right]^2.
\]

Working along similar lines, an approximate \((1 - \alpha)\)-level of CI for specificity by Woodroofe’s pivot method can be shown to be

\[
\left\{ \frac{S^-_M}{n^-_M} - \frac{\hat{\delta}(sp)}{\sqrt{n^-_M}} + \frac{\sigma(sp)}{\sqrt{n^-_M}} Z_{\alpha/2} \right\}.
\]

(4.3)

(b) Coverage probabilities and expected length of CIs

Because the stopping boundaries in sequential designs are constructed on a two-dimensional test statistic, CIs for sensitivity and specificity in equations (4.2) and (4.3), respectively, cannot be expressed by a single parameter.

Tables 1 and 2 show the coverage probability and the expected length of the CIs for sensitivity and specificity, respectively, based on a simulation study of 5000 iterations for five optimal two-stage designs. As defined in §3, \(a^+_1\) and \(a^-_1\) are the stopping boundaries in stage 1, \(n^+_1\) and \(n^-_1\) are the required number of diseased and non-diseased subjects, and \(n_2\) is the number of subjects per group at the end of stage 2 with equal allocation between two groups. For each two-stage design, three sets of the underlying true values of parameters are specified in the second column. The CIs constructed by Woodrooffe’s pivot method are based on different plug-in estimators, i.e. W.MLE for the maximum-likelihood estimators and W.RB for the Rao–Blackwell unbiased estimators. These CIs are compared with the conventional \(100(1 - \alpha)\%\) CIs constructed based on the maximum-likelihood estimates, i.e.

\[
\text{CI}_{\text{MLE}} : \hat{\theta}_{\text{MLE}_i} \pm Z_{\alpha/2} \sqrt{\text{Var}(\hat{\theta}_{\text{MLE}_i})}, \quad i = 1, 2.
\]

(4.4)
Table 1. Coverage probability and expected length of 95% CIs for sensitivity in the optimal two-stage designs.

<table>
<thead>
<tr>
<th>design $(a^+_1, a^-_1)/(n^+_1, n^-_1)/n_2$</th>
<th>parameters (se, sp)</th>
<th>CP$_{MLE}$</th>
<th>CP$_{W,MLE}$</th>
<th>CP$_{W,RB}$</th>
<th>EL$_{MLE}$</th>
<th>EL$_{W,MLE}$</th>
<th>EL$_{W,RB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5, 7)/(8, 10)/21</td>
<td>(0.70, 0.75)</td>
<td>0.949</td>
<td>0.948</td>
<td>0.920</td>
<td>0.570</td>
<td>0.547</td>
<td>0.540</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.920</td>
<td>0.920</td>
<td>0.892</td>
<td>0.486</td>
<td>0.471</td>
<td>0.407</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.894</td>
<td>0.892</td>
<td>0.892</td>
<td>0.321</td>
<td>0.264</td>
<td>0.259</td>
</tr>
<tr>
<td>(10, 10)/(14, 13)/26</td>
<td>(0.70, 0.75)</td>
<td>0.930</td>
<td>0.911</td>
<td>0.911</td>
<td>0.452</td>
<td>0.452</td>
<td>0.449</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.869</td>
<td>0.912</td>
<td>0.912</td>
<td>0.383</td>
<td>0.359</td>
<td>0.353</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.926</td>
<td>0.925</td>
<td>0.922</td>
<td>0.262</td>
<td>0.232</td>
<td>0.229</td>
</tr>
<tr>
<td>(10, 9)/(14, 12)/32</td>
<td>(0.70, 0.75)</td>
<td>0.931</td>
<td>0.923</td>
<td>0.923</td>
<td>0.450</td>
<td>0.450</td>
<td>0.447</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.902</td>
<td>0.910</td>
<td>0.910</td>
<td>0.381</td>
<td>0.350</td>
<td>0.343</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.821</td>
<td>0.820</td>
<td>0.816</td>
<td>0.251</td>
<td>0.215</td>
<td>0.213</td>
</tr>
<tr>
<td>(4, 5)/(6, 7)/22</td>
<td>(0.70, 0.75)</td>
<td>0.918</td>
<td>0.897</td>
<td>0.897</td>
<td>0.637</td>
<td>0.612</td>
<td>0.600</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.878</td>
<td>0.876</td>
<td>0.876</td>
<td>0.558</td>
<td>0.483</td>
<td>0.462</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.876</td>
<td>0.875</td>
<td>0.870</td>
<td>0.400</td>
<td>0.302</td>
<td>0.284</td>
</tr>
<tr>
<td>(8, 7)/(11, 9)/28</td>
<td>(0.70, 0.75)</td>
<td>0.896</td>
<td>0.886</td>
<td>0.886</td>
<td>0.499</td>
<td>0.500</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.920</td>
<td>0.919</td>
<td>0.918</td>
<td>0.432</td>
<td>0.404</td>
<td>0.395</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.926</td>
<td>0.924</td>
<td>0.929</td>
<td>0.296</td>
<td>0.243</td>
<td>0.236</td>
</tr>
</tbody>
</table>

$^a$CP$_{MLE}$ and EL$_{MLE}$ are computed based on the naive approximate CI in equation (4.4); CP$_{W,MLE}$, EL$_{W,MLE}$ and CP$_{W,RB}$, EL$_{W,RB}$ are based on equation (4.2) using maximum-likelihood and Rao–Blackwell estimators, respectively.

Table 2. Coverage probability and expected length of 95% CIs for specificity in the optimal two-stage designs.

<table>
<thead>
<tr>
<th>design $(a^+_1, a^-_1)/(n^+_1, n^-_1)/n_2$</th>
<th>parameters (se, sp)</th>
<th>CP$_{MLE}$</th>
<th>CP$_{W,MLE}$</th>
<th>CP$_{W,RB}$</th>
<th>EL$_{MLE}$</th>
<th>EL$_{W,MLE}$</th>
<th>EL$_{W,RB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5, 7)/(8, 10)/21</td>
<td>(0.70, 0.75)</td>
<td>0.948</td>
<td>0.915</td>
<td>0.915</td>
<td>0.475</td>
<td>0.472</td>
<td>0.464</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.810</td>
<td>0.810</td>
<td>0.808</td>
<td>0.368</td>
<td>0.342</td>
<td>0.335</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.646</td>
<td>0.646</td>
<td>0.645</td>
<td>0.173</td>
<td>0.155</td>
<td>0.154</td>
</tr>
<tr>
<td>(10, 10)/(14, 13)/26</td>
<td>(0.70, 0.75)</td>
<td>0.888</td>
<td>0.873</td>
<td>0.873</td>
<td>0.426</td>
<td>0.438</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.895</td>
<td>0.894</td>
<td>0.892</td>
<td>0.338</td>
<td>0.325</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.732</td>
<td>0.731</td>
<td>0.729</td>
<td>0.169</td>
<td>0.152</td>
<td>0.150</td>
</tr>
<tr>
<td>(10, 9)/(14, 12)/32</td>
<td>(0.70, 0.75)</td>
<td>0.876</td>
<td>0.860</td>
<td>0.860</td>
<td>0.441</td>
<td>0.453</td>
<td>0.449</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.938</td>
<td>0.937</td>
<td>0.937</td>
<td>0.350</td>
<td>0.323</td>
<td>0.318</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.789</td>
<td>0.788</td>
<td>0.786</td>
<td>0.167</td>
<td>0.144</td>
<td>0.142</td>
</tr>
<tr>
<td>(4, 5)/(6, 7)/22</td>
<td>(0.70, 0.75)</td>
<td>0.904</td>
<td>0.888</td>
<td>0.888</td>
<td>0.548</td>
<td>0.540</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.853</td>
<td>0.774</td>
<td>0.770</td>
<td>0.433</td>
<td>0.391</td>
<td>0.373</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.621</td>
<td>0.621</td>
<td>0.620</td>
<td>0.213</td>
<td>0.173</td>
<td>0.166</td>
</tr>
<tr>
<td>(8, 7)/(11, 9)/28</td>
<td>(0.70, 0.75)</td>
<td>0.914</td>
<td>0.924</td>
<td>0.905</td>
<td>0.494</td>
<td>0.507</td>
<td>0.501</td>
</tr>
<tr>
<td></td>
<td>(0.80, 0.85)</td>
<td>0.869</td>
<td>0.869</td>
<td>0.862</td>
<td>0.399</td>
<td>0.382</td>
<td>0.367</td>
</tr>
<tr>
<td></td>
<td>(0.90, 0.95)</td>
<td>0.735</td>
<td>0.734</td>
<td>0.731</td>
<td>0.211</td>
<td>0.172</td>
<td>0.162</td>
</tr>
</tbody>
</table>

$^a$CP$_{MLE}$ and EL$_{MLE}$ are computed based on the naive approximate CI in equation (4.4); CP$_{W,MLE}$, EL$_{W,MLE}$ and CP$_{W,RB}$, EL$_{W,RB}$ are based on equation (4.3) using maximum-likelihood and Rao–Blackwell estimators, respectively.
It is observed that the coverage probabilities of the CIs in the simulation study are under the nominal level for small sample sizes in tables 1 and 2. When the true values of the parameters are closer to 1.0, the coverage probabilities get lower. Approximating a binomial distribution with a normal distribution becomes inaccurate when the binomial parameter $p$ is close to points 0 and 1. In spite of less coverage probabilities than the nominal level, the CIs by Woodroofe’s pivot method using the maximum-likelihood and the Rao–Blackwell estimators tend to have shorter interval length with competitive coverage probabilities than the normal approximated intervals based on the maximum-likelihood estimates.

5. Discussion

We evaluated the bias and mean squared error of the point estimates of sensitivity and specificity in group sequential designs, and demonstrated the difference between maximum-likelihood and Rao–Blackwell unbiased estimates using an optimal two-stage design proposed by Shu et al. (2007). Although the magnitude of the bias of the maximum-likelihood estimates is relatively less substantial in two-stage designs, it is expected to increase if the number of interim analyses increases.

Whitehead (1986) suggests adjusting the maximum-likelihood estimate by subtracting an estimate of the bias of the MLE. Another approach, suggested by Wang & Leung (1997), is to use parametric bootstrap methods to find bias-adjusted estimators. The applications of Whitehead’s bias correction method and Wang & Leung’s bootstrap methods in the group sequential designs of sensitivity and specificity are of interest for future studies.

The simulation results for the CIs show that the normal approximation approach does not work well in small samples. The coverage probabilities are less than the nominal level of confidence, especially when the estimates are closer to the boundaries. One approach is to construct the CI based on a transformation of the parameters. A logit transformation is a traditional choice for estimating proportions. Furthermore, Whitehead et al. (2000) described methods for setting CIs for secondary parameters in a way that provides the correct coverage probability in repeated frequentist realizations of the sequential design used.

The challenge of constructing exact CIs of the parameters is how to order the two-dimensional sample space following a multistage test. There are a variety of ways to order sample points on stopping boundaries and across sequential stages. With a specific ordering, a more ‘extreme’ case can be defined and exact CIs can be computed. Further research is much needed to develop these methods.

References


Phil. Trans. R. Soc. A (2008)
Liu, A., Wu, C., Yu, K. F. & Gehan, E. 2005 Supplementary analysis of probabilities at
sim.1990)
Liu, A., Hall, W. J., Yu, K. F. & Wu, C. 2006 Estimation following a group sequential test for
Obuchowski, N. A. & Zhou, X. H. 2002 Prospective studies of diagnostic test accuracy when
Shu, Y., Liu, A. & Li, Z. 2007 Sequential evaluation of a medical diagnostic test with binary
Todd, S. & Whitehead, J. 1997 Confidence interval calculation for a sequential clinical trial of
binary response. Biometrika 84, 737–743. (doi:10.1093/biomet/84.3.737)
Wang, Y. G. & Leung, D. H. Y. 1997 Bias reduction via resampling for estimation following
sequential tests. Sequential Anal. 16, 249–267. (doi:10.1080/07474949708836386)
Whitehead, J. 1986 On the bias of maximum likelihood estimation following a sequential test.
Biometrika 73, 573–581. (doi:10.1093/biomet/73.3.573)
Whitehead, J., Todd, S. & Hall, W. J. 2000 Confidence intervals for secondary parameters
1467-9868.00260)
Woodroofe, M. 1992 Estimation after sequential testing: a simple approach for a truncated
Wruck, L. M., Yiannoutsos, C. T. & Hughes, M. D. 2006 A sequential design to estimate sensitivity