Published in final edited form as:

Stat Med. 2016 August 30; 35(19): 3397–3412. doi:10.1002/sim.6932.

Repeated significance tests of linear combinations of sensitivity and specificity of a diagnostic biomarker

Mixia Wu¹, Yu Shu², Zhaohai Li³, and Aiyi Liu^{4,*†}

¹College of Applied Sciences, Beijing University of Technology, Beijing 100124, China

²Abbott Labs Vascular Division, 3200 Lakeside Dr., Santa Clara, CA 95054, U.S.A

³Department of Statistics, George Washington University, 801 22nd Street, Washington DC, 20037, U.S.A

⁴Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Blvd., Rockville, MD 20852, U.S.A

Abstract

A sequential design is proposed to test whether the accuracy of a binary diagnostic biomarker meets the minimal level of acceptance. The accuracy of a binary diagnostic biomarker is a linear combination of the marker's sensitivity and specificity. The objective of the sequential method is to minimize the maximum expected sample size under the null hypothesis that the marker's accuracy is below the minimal level of acceptance. The exact results of two-stage designs based on Youden's index and efficiency indicate that the maximum expected sample sizes are smaller than the sample sizes of the fixed designs. Exact methods are also developed for estimation, confidence interval and p-value concerning the proposed accuracy index upon termination of the sequential testing.

Keywords

Expected sample size; inference upon early stopping; sensitivity; sequential testing; specificity; two-stage designs; type I error; Youden's index

1. Introduction

In evaluating the performance of a diagnostic biomarker (or test), the sensitivity (se) and specificity (sp) are important accuracy measures. The sensitivity is the probability that the biomarker correctly classifies a true case of the disease as diseased, and the specificity is the probability that it correctly classifies a non-case as non-diseased [1–2]. Sensitivity and specificity are often estimated for the new diagnostic biomarker applied to independent random samples of patients with known disease status. Besides sensitivity and specificity,

*Correspondence to: Aiyi Liu, Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Rockville, MD 20852, U.S.A.

†liua@mail.nih.gov

the accuracy of a diagnostic biomarker can be defined in various ways. Youden [3] proposed a one-dimension summary of classification, known as the Youden's index.

Youden's index gives equal weights to sensitivity and specificity, which exploits the information provided by the diagnostic biomarker in the context of a particular diagnostic objective, and facilitate comparison of different diagnostic biomarkers [4]. The advantage of Youden's index is that it does not require knowledge of the prevalence of a disease. The sum of sensitivity and specificity is also used to evaluate whether a binary decision rule operates better than chance, and an exact confidence interval for the accuracy of the biomarker was provided by Parker and Davis [5]. However, a drawback of such a one-dimension measure is that it cannot alone capture how a biomarker performs in both disease and disease-free populations [6]. With the knowledge of the prevalence (π) of the condition in the target population, other measures for accuracy are also expressed in term of sensitivity and specificity. Efficiency ($Eff = \pi se + (1 - \pi)sp$) and a misclassification-cost term ($MCT = (C_{FN}/C_{FP})\pi(1 - se) + (1 - \pi)(1 - sp)$, where C_{FP} and C_{FN} represent the utilities associated with the two false test outcomes [7], are good examples of linear combinations of sensitivity and specificity with weights selected to reflect the decision-making situation. Predictive values are the accuracies that can be quantified by how well the test result predicts true disease status [8]. The positive predictive value ($PPV = \pi se / \{\pi se + (1 - \pi)(1 - sp)\}$) is the probability that the condition is present given the biomarker indicates its presence; and the negative predictive value ($NPV = (1 - \pi)sp / \{\pi(1 - se) + (1 - \pi)sp\}$) is the probability that the condition is absent given the biomarker indicates its absence. Both measures are non-linear combinations of the sensitivity and specificity, and so is the odds ratio, $OR = (se/(1 - se))/(sp/(1 - sp))$, which is independent of the prevalence rate.

We consider an index J as a linear combination of the sensitivity and specificity with various weights:

$$J = \lambda se + \mu sp - \phi, \quad (1)$$

where λ , μ and ϕ are known positive coefficients, usually chosen so that $\max_{\{se, sp\}} J = \lambda + \mu - \phi > 0$ for practical use. Both sensitivity and specificity are special cases of the index. Furthermore, when $\lambda = \mu = \phi = 1$, J reduces to the popular Youden's index, and when $\lambda = \pi$, $\mu = 1 - \pi$ and $\phi = 0$, the index becomes efficiency. The combination coefficients, λ and μ , can also be considered as the potential risk associated with the false positive fraction and false negative fraction in order to balance the cost of a study.

In Phase II development of a diagnostic biomarker, as discussed in [8], the objective of the study is to determine whether a new diagnostic biomarker meets the minimal level of acceptance for accuracy and formal sample size calculation is typically required in the study. Human tissues or blood samples are often used in many diagnostic studies; some procedures can result in discomfort in patients, and can be expensive in cost. Moreover, in the exploratory phases of the development process, multiple diagnostic biomarkers exist and are typically compared to a minimal acceptable level of accuracy in a case-control study with limited number of subjects from diseased and non-diseased populations. In such a case, there

is a need to stop evaluation of ineffective biomarkers early so that limited resources can be reallocated to evaluate more promising biomarkers. All these stated scenarios call for early termination of the study and minimization of the number of required samples when evidence indeed shows the diagnostic biomarker under investigation is inefficient in classification.

To this end, group sequential methods that have been well-developed and widely used in clinical trial area appear to be most appealing in addressing these concerns stated above. Well-detailed in [9–10], among others, a group sequential procedure tests a null hypothesis a few times using accumulated data while controlling the overall type I error rate. Early termination of the trial may occur when a decision can be made about rejection of the null hypothesis. In large-scale comparative phase III trial a group sequential procedure is usually carried out based on large sample normal approximation. For small-scale early phase trials with binary outcomes, exact calculations based on binomial theory are quite tedious and were developed mainly for two-stage phase II trials with a single or bivariate binary outcomes; see [11–14]. Normal theory based group sequential methods have been extended to the field of diagnostic medicine to compare the areas of continuous biomarkers under their receiver operating characteristic curves [15–17]. For simultaneous testing of hypotheses concerning both sensitivity and specificity of a binary biomarker, a two-stage procedure with exact calculations was developed in [18].

When the index (1) is used as the accuracy measure, we propose along these lines a group sequential (two-stage) design for testing hypotheses concerning the index, allowing early termination of the evaluation process if the index falls below the minimal acceptance level. Since the focus is early stage evaluation of diagnostic biomarkers, large sample theory and asymptotic normal approximation will not work satisfactorily. Therefore, in the present paper, the binomial probability distribution is used to obtain the exact optimal sequential designs.

The remainder of this paper is organized as follows. In Section 2, we start with a proposed fixed sample size design, followed by a sequential procedure that minimizes the maximum expected sample size under the null hypothesis, and provide with explicit details of two-stage design in Section 3. In Section 4, Youden's index is used as a special case to illustrate that the expected sample sizes of two-stage designs are smaller than the sample sizes required by the fixed-size designs. Similar patterns are observed in the example of efficiency. Numerical results of a number of two-stage designs are presented as well. Normal approximation method is discussed and compared to the proposed optimal multi-stage designs in Section 5. Section 6 proposes exact methods for terminal inference on the index, including estimation, confidence interval and p-value. We illustrate our method with an example of HIV-1 infection classification in Section 7. Finally, the strength and limitation of the methods, along with some comments on future research, are discussed in Section 8.

2. Fixed sample size design

Suppose that the investigators want to evaluate a diagnostic biomarker based on a linear combination J of its sensitivity and specificity. Random samples of n^D cases and n^D controls are selected independently from the diseased and disease-free populations. The null

hypothesis, H_0 , is that J is no larger than a pre-specified number θ_0 , herewithin referred to as the minimal level of acceptance. In practice, this minimal level of acceptance can be obtained from empirical data on similar existing diagnostic biomarkers, or based on the minimal level of acceptance for sensitivity se_0 and specificity sp_0 , resulting in $\theta_0 = \lambda se_0 + \mu sp_0 - \phi$. The alternative hypothesis is $H_1 : J > \theta_0$. The null hypothesis is rejected when a plausible estimate $\hat{J}$ of J exceeds a critical value $b \in (-\phi, \lambda + \mu - \phi)$. In a fixed-size design, let S^D and $S^{\bar{D}}$ be respectively the number of correctly classified diseased and non-diseased samples, then J is estimated by $\lambda \frac{S^D}{n^D} + \mu \frac{S^{\bar{D}}}{n^{\bar{D}}} - \phi$, the maximum likelihood estimate of J . The power function of the test can be expressed as a function of any pair of sensitivity, specificity, and $J = \theta > \theta_0$. Given sensitivity se and $J = \theta$, the specificity is $(\theta + \phi - \lambda se)/\mu$. From the binomial theory, the power function has the following expression:

$$\begin{aligned}\beta(se, \theta) &= P(\hat{J} > b) = P\left(\lambda \frac{S^D}{n^D} + \mu \frac{S^{\bar{D}}}{n^{\bar{D}}} - \phi > b\right) \\ &= \sum_i \sum_j d(i, n^D, se) d(j, n^{\bar{D}}, (\theta + \phi - \lambda se)/\mu),\end{aligned}$$

where $d(\cdot, \cdot, \cdot)$ is the probability mass function of the binomial distribution, with the last argument being the binomial probability. The i and j are summed over

$$\begin{aligned}i &\in \left[\left((b + \phi - \mu) \frac{n^D}{\lambda}, n^D \right], \\ j &\in \left[\left((b + \phi - \lambda \frac{i}{n^D}) \frac{n^{\bar{D}}}{\mu}, n^{\bar{D}} \right].\end{aligned}$$

For every $\theta > \theta_0$, the probability of making the type I error, $\beta(se, \theta)$, is a function of the parameter se , and is nondecreasing in θ . It thus suffices to consider controlling $\beta(se, \theta_0)$. The maximum value of $\beta(se, \theta_0)$ over the possible range of se should be controlled no more than α , while under the alternative hypothesis, the minimum power should be no less than $1 - \beta$. In the alternative parameter space when $J = \theta_1$, the investigators are commonly interested in those diagnostic biomarkers whose sensitivity and specificity are both at high level. In order to compare the power of a fixed-size design and that of a sequential design, we only consider the required power $1 - \beta$ at a specific point when sensitivity is at se_1 , and specificity at $\frac{\theta_1 + \phi - \lambda se_1}{\mu}$. Then, the critical value b can be obtained by solving the following equations with the specified value of α and β :

$$\alpha = \max_{se} \beta(se, \theta_0), \quad (2)$$

$$1 - \beta = \beta(se_1, \theta_1). \quad (3)$$

In the present paper, the maximization in (2) is carried out over the possible range of se that is,

$$\max\{0, (\theta_0 + \phi - \mu)/\lambda\} \leq se \leq \min\{1, (\theta_0 + \phi)/\lambda\}.$$

In practice, however, a narrower range can be applied if prior knowledge exists about the sensitivity (and or specificity).

Based on a discrete probability function, there is no closed form for sample size calculation. A numerical search procedure is developed to obtain the sample sizes that satisfies type I error and power constraints. With known allocation ratio, $n^D/n^{\bar{D}} = \eta$, the procedure begins with a small initial sample size of n^D cases and $n^{\bar{D}}$ controls. The largest possible value of critical value b is determined such that the power of the test is no less than $1 - \beta$. Once se_1 and θ_1 are known, the power function is a non-increasing function of the critical value b . Based on n^D , $n^{\bar{D}}$ and b , we check whether the maximum probability of type I error is no more than α . The maximum probability of type I error under the null hypothesis can be computed by various optimizing methods. If the type I error rate constraint is not satisfied, then we increase the sample sizes with the same allocation ratio η and continue the searching procedure. Multiple designs satisfying both constraints may exist with an increase in the sample sizes n^D and $n^{\bar{D}}$. As the criteria of selection is to keep the sample sizes as low as possible, the final fixed-size design is chosen to be the one with the minimal total sample sizes satisfying the significance level and power.

3. Sequential designs

We focus on assessing the performance of a diagnostic biomarker based on the minimal acceptance level of a linear combination of sensitivity and specificity. If the evidence from the early stage shows the accuracy of the diagnostic biomarker to be statistically below the required acceptable level, then the study should be terminated without further investigation. Suppose that investigators decide to perform a total of K interim analyses to test the same hypotheses. At the end of the k^{th} stage, $k = 1, 2, \dots, K - 1$, a total number of n_k^D cases and $n_k^{\bar{D}}$ controls are tested, among which S_k^D and $S_k^{\bar{D}}$ are classified as diseased and non-diseased, respectively. Let $\hat{J}_k = \lambda S_k^D / n_k^D + \mu S_k^{\bar{D}} / n_k^{\bar{D}} - \phi$ be the estimate of J at stage k . If $\hat{J}_k > b_k$, then the study continues to stage $k + 1$; otherwise it is terminated because of the diagnostic biomarker's low accuracy. If the study does not stop at early stage $k = 1, 2, \dots, K - 1$, then at stage K , reject the null hypothesis if $\hat{J}_K > b_K$, and the diagnostic biomarker will undergo a large scale of study, often a comparative study to compare it with other diagnostic biomarkers. Note that these types of sequential testing stop early only for futility of the diagnostic biomarker; the null hypothesis could be rejected only at the last stage.

3.1. Two-stage design method

With small number of subjects available or tight budget, it is practically difficult to have a design with many interim analyses. Two-stage design is quick and efficient to save resources when there are other candidate biomarkers needed to be evaluated. Moreover, with the increasing number of interim analyses, the power function becomes complicated and the numerical computation can be time consuming.

In a two-stage design, the probability of terminating the study and rejecting the diagnostic biomarker is:

$$P(\text{Reject the diagnostic biomarker}) = P(\hat{J}_1 \leq b_1) + P(\hat{J}_2 \leq b_2, \hat{J}_1 > b_1). \quad (4)$$

Then, the probability of rejecting H_0 at θ is:

$$1 - P(\text{Reject the diagnostic biomarker}) = P(\hat{J}_2 > b_2, \hat{J}_1 > b_1). \quad (5)$$

Similar to the fixed size design, the power function in equation (5) depends on not only $J = \theta$ but also sensitivity (or specificity). Indeed, utilizing the binomial theory and the fact that

$(S_1^D, S_1^{\bar{D}})$ and $(S_2^D - S_1^D, S_2^{\bar{D}} - S_1^{\bar{D}})$ are independent, it can be expressed as:

$$\begin{aligned} \beta(se, \theta) &= P(\hat{J}_1 > b_1, \hat{J}_2 > b_2) = P\left(\lambda \frac{S_1^D}{n_1^D} + \mu \frac{S_1^{\bar{D}}}{n_1^{\bar{D}}} - \phi > b_1, \lambda \frac{S_2^D}{n_2^D} + \mu \frac{S_2^{\bar{D}}}{n_2^{\bar{D}}} - \phi > b_2\right) \\ &= \sum_i \sum_j \sum_k \sum_l d(i, n_1^D, se) d(j, n_1^{\bar{D}}, \frac{\theta + \phi - \lambda se}{\mu}) d(k, n_2^D - n_1^D, se) \times d(l, n_2^{\bar{D}} - n_1^{\bar{D}}, \frac{\theta + \phi - \lambda se}{\mu}), \end{aligned}$$

where the summations are over

$$\begin{aligned} i &\in \left((b_1 + \phi - \mu) \frac{n_1^D}{\lambda}, n_1^D \right], \\ j &\in \left((b_1 + \phi - \lambda \frac{i}{n_1^{\bar{D}}}) \frac{n_1^{\bar{D}}}{\mu}, n_1^{\bar{D}} \right], \\ k &\in \left(\left(b_2 + \phi - (n_2^{\bar{D}} - n_1^{\bar{D}} + j) \frac{\mu}{n_2^{\bar{D}}} \right) \frac{n_2^D}{\lambda} - i, n_2^D - n_1^D \right], \\ l &\in \left(\left(b_2 + \phi - \lambda \frac{(i+k)}{n_2^{\bar{D}}} \right) \frac{n_2^{\bar{D}}}{\mu} - j, n_2^{\bar{D}} - n_1^{\bar{D}} \right]. \end{aligned}$$

The optimal two-stage design needs to satisfy the following two requirements:

$$\alpha = \max_{se} \beta(se, \theta_0), \quad (6)$$

$$1 - \beta = \beta(se_1, \theta_1). \quad (7)$$

The expected sample size, also called the average sample number/size, is an important characteristic of any sequential procedures. As stated in the introduction, the expected number of subjects to participate a diagnostic biomarker needs to be minimized in a sequential design for the budget or ethical considerations. For a two-stage design stated above, let $PST_{se, \theta}$ be the probability of stopping the biomarker at stage 1, that is

$$PST_{se,\theta} = P(\text{Stopping at stage 1}) = P(\hat{J}_1 \leq b_1). \quad (8)$$

Then its expected sample size is given by

$$\begin{aligned} ASN_{se,\theta} &= (n_1^D + n_1^{\overline{D}})P(\text{Stopping at stage 1}) + (n_2^D + n_2^{\overline{D}})P(\text{Stopping at stage 2}) \\ &= (n_1^D + n_1^{\overline{D}})PST_{se,\theta} + (n_2^D + n_2^{\overline{D}})(1 - PST_{se,\theta}). \end{aligned} \quad (9)$$

The expected sample size is a function of $J = \theta$ and se through $\lambda se + \mu sp - \varphi = \theta$. Since θ is the primary parameter of interest, to remove the nuisance parameter se , we consider the maximum expected sample size among all possible values of se

$$ASN_{\theta}^* = \max_{se} ASN_{se,\theta}. \quad (10)$$

Note that both $PST_{se,\theta}$ and ASN_{θ}^* depend also on the corresponding sample sizes. Our aim is to find a design that minimizes the maximum expected sample size. With pre-specified value of $(\alpha, \beta, \theta_0, \theta_1, se_1)$, each two-stage design is characterized by (n_1, b_1, n_2, b_2) . There are many two-stage designs that satisfy error constraints (6) and (7), and there are different ways to define an optimal two-stage design. A number of optimal criteria for two-stage designs have been discussed in phase II clinical trials; see, among others, [11–14]. In the present paper, a two-stage design satisfying (6) and (7) that minimizes the maximum expected sample number under the null hypothesis ($J = \theta_0$) is referred to as the optimal two-stage design.

3.2. Numerical algorithm

In the algorithm of sample sizes determination, we assume that the allocation of the cases and controls in each stage is determined by the values of λ and μ , i.e., $\lambda/\mu = n_k^D/n_k^{\overline{D}}$, $k = 1, 2, \dots, K$. For example, Youden's index is a proper measure when costs associated with the two types of false results are similar in magnitude, in which case it is reasonable to have equal allocation between the cases and the controls. In order to compare the sample sizes of two-stage designs and those of fixed size designs, $J = \theta_1$ and the sensitivity se_1 under the alternative hypothesis are to be specified for the required power. The implementation of the algorithm is similar when, instead of sensitivity, the specificity sp_1 under the alternative hypothesis is specified.

The algorithm numerically searches for the optimal two-stage design in order to minimize the maximum expected sample size $ASN_0^* = ASN_{\theta_0}^*$ at $J = \theta_0$. Under known allocation

assumption, for fixed sample sizes of cases and controls $\left(\frac{n_2^D}{n_2^{\overline{D}}} = \frac{\lambda}{\mu} = \eta\right)$, for each pair of n_1^D and $n_1^{\overline{D}}$ with known allocation ratio, the algorithm lists all possible critical values

$b_1 = S_1^D / n_1^D + S_1^{\overline{D}} / n_1^{\overline{D}} - 1$, where $S_1^D = 0, 1, \dots, n_1^D$ and $S_1^{\overline{D}} = 0, 1, \dots, n_1^{\overline{D}}$. Then, for each distinguishing value of b_1 , the algorithm determines the largest critical value in the second stage b_2 satisfying the required power. Based on a pair of the critical values (b_1, b_2) satisfying the power, the maximum probability of type I error is computed at $J = \theta_0$. If the maximum type I error rate, $\max_{se} \beta(se, \theta_0)$, is no more than the nominal significance level α , the maximum expected sample size based on this particular design is obtained and compared with the previous feasible design. If the minimal maximum expected sample size is achieved by multiple feasible two-stage designs for its fixed total sample size pair $(n_2^D, n_2^{\overline{D}})$, we choose the one with the smallest difference to the nominal significance level. Next, the algorithm starts over for a new search when increasing the total sample size, and preserving the allocation ratio. The minimal maximum expected sample size under H_0 is not monotonic function of the total sample size $n_2^D + n_2^{\overline{D}}$. However, with the increase in the total sample size, the global minimal will be achieved. The maximum values of the power function and the expected sample size under H_0 are obtained by optimization procedures.

4. Numerical results and simulation study

4.1. Numerical results

Youden's index is a special case of J when all coefficients λ , μ and φ equal to 1. In Table 1, we present the two-stage designs for a variety of design parameters for testing Youden's index. Since $\lambda = \mu = 1$ in Youden's index, we used equal allocation of the cases and controls for both stages, i.e., $n_k^D = n_k^{\overline{D}} = n_k$, $k = 1, 2$, so that n_1 and n_2 are the required sample size at the end of the first and second stage per group. Denote by n_f the required sample size per group at the point of se_1 and $sp_1 = \theta_1 + 1 - se_1$ under the alternative hypothesis in a fixed size design with the same level of power. In Table 1, we chose sensitivity se_1 under the alternative parameter space to be $\frac{\theta_1 + 1}{2}$. The theoretical probability of type I error α^* and power $1 - \beta^*$ achieved by the optimal two-stage design are also presented in the table.

The results in Table 1 show that the required two-stage sample sizes per group (n_2) are no less than that of the fixed size per group (n_f). However, as expected the maximum expected sample size ASN_0^* under H_0 for the two-stage designs are much smaller than the total sample sizes for the fixed size designs $2 \times n_f$. With a few extra patients in each group, the optimal two-stage designs allow investigators to terminate the study earlier and minimize the expected number of patients for diagnostic biomarkers having poor accuracy.

We also demonstrate the proposed method in Table 2 with an example of efficiency, assuming that the disease has a prevalence of 25 per cent in the population. Thus

$Eff = \frac{1}{4}se + \frac{3}{4}sp$, $\lambda: \mu = n_k^D : n_k^{\overline{D}} = 1:3$ and $\varphi = 0$. The efficiency can be interpreted as the proportion of subjects in the population classified by the diagnostic biomarker as diseased. The results in Table 2 are consistent with the conclusion from Table 1, that the maximum expected sample sizes of the optimal two-stage designs are less than the sample sizes required by the fixed size designs with the same desired power.

4.2. Simulation studies

To validate the proposed method and numerical algorithm, we conducted simulation study to compare the theoretical significance level and the maximum expected sample sizes under H_0 for the optimal two-stage designs to the simulation results.

In order to generate random variables under binomial distributions, both sensitivity and specificity need to be specified. Over a fine grid of sensitivity and specificity in the parameter space of H_0 , for each pair of fixed level of sensitivity and specificity, a single value of the significance level is obtained by a large number of iterations. For each iteration, in stage 1, the number of positive results S_1^D is generated from n_1^D cases and the number of negative results $S_1^{\bar{D}}$ is generated from $n_1^{\bar{D}}$ controls. If the test statistic J_1 exceeds the critical value b_1 , then another pair of binomial random variables ($X_2^D, X_2^{\bar{D}}$) are generated from $n_2^D - n_1^D$ cases and $n_2^{\bar{D}} - n_1^{\bar{D}}$ controls in the second stage, respectively. Subsequently, $\hat{J}_2 = \lambda \frac{S_1^D + X_2^D}{n_2^D} + \mu \frac{S_1^{\bar{D}} + X_2^{\bar{D}}}{n_2^{\bar{D}}} - \phi$ is compared to the critical value b_2 to determine whether or not to reject the null hypothesis. The maximum significance level is obtained by repeating this process for a equal space grid over parameter space under H_0 .

Table 3 and Table 4 illustrate the simulation results corresponding to the designs listed in Table 1 and Table 2, respectively. In the table α_s and ASN_s are simulated type I error rate and the maximum expected sample size based on 100,000 iterations. Again, α^* and ASN_0^* are the theoretical values of the optimal two-stage designs.

The results show that the simulated probability of type I error rates α_s appear to be slightly larger than the theoretical type I error rates α^* in most cases. The relatively small sample sizes in the first stage contribute to the difference. Additionally, the critical values are rounded in the simulation procedure, so the less precision in the critical values in both stages can also result in the difference between theoretical and simulated results. The simulated and theoretical maximum expected sample sizes are close.

5. Normal approximation methods of interim analyses of index J

When sample size is sufficiently large, the sample proportion has approximate normal distribution. Sensitivity and specificity in case-control studies are the parameters of two independent binomial distribution. In this section, we discuss a normal approximation method of testing the index J .

Consider a group sequential study with a maximum of K group of observations. Let $\beta^{(k)}$, $k = 1, \dots, K$, denote the estimates of parameter vector $\beta = (se, sp)^T$ at successive analyses. The maximum likelihood estimator $(\hat{\beta}^{(1)}, \dots, \hat{\beta}^{(K)})$, $\hat{\beta}^{(k)} = (S_k^D/n_k^D, S_k^{\bar{D}}/n_k^{\bar{D}})^T$ is approximately multivariate normal with

$$\hat{\boldsymbol{\beta}}^{(k)} \sim N(\boldsymbol{\beta}, \mathcal{I}^{-1}(k, \boldsymbol{\beta})), \quad k=1, \dots, K,$$

$$\text{Cov}(\hat{\boldsymbol{\beta}}^{(k_1)}, \hat{\boldsymbol{\beta}}^{(k_2)}) = \text{Var}(\hat{\boldsymbol{\beta}}^{(k_2)}) = \mathcal{I}^{-1}(k_2, \boldsymbol{\beta}), \quad k_1 \leq k_2.$$

$\mathcal{I}^{-1}(k, \boldsymbol{\beta})$, $k = 1, \dots, K$ is 2×2 matrix

$$\begin{pmatrix} se(1-se)/n_k^D & 0 \\ 0 & sp(1-sp)/n_k^{\overline{D}} \end{pmatrix},$$

Suppose we wish to test the hypothesis $H_0: \mathbf{c}^T \boldsymbol{\beta} = \gamma$ for a given vector $\mathbf{c}$ and scalar constant γ . Define $\theta = \mathbf{c}^T \boldsymbol{\beta} - \gamma$ and

$$\mathcal{I}_k = \{ \text{Var}(\mathbf{c}^T \hat{\boldsymbol{\beta}}^{(k)}) \}^{-1} = (\mathbf{c}^T \mathcal{I}^{-1}(k, \boldsymbol{\beta}) \mathbf{c})^{-1}, \quad (11)$$

the Fisher information for $\mathbf{c}^T \boldsymbol{\beta}$ at analysis k . Let $\hat{\mathcal{I}}_k$ be an estimate of $\mathcal{I}_k$ by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}^{(k)}$ in (11). Then a test can be defined in terms of the Wald statistics,

$$Z_k = (\mathbf{c}^T \hat{\boldsymbol{\beta}}^{(k)} - \gamma) \sqrt{\hat{\mathcal{I}}_k}, \quad k=1, \dots, K, \quad (12)$$

which are jointly multivariate normal with

$$\begin{cases} Z_k \sim N(\theta \sqrt{\mathcal{I}_k}, 1), & k=1, \dots, K, \\ \text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1} / \mathcal{I}_{k_2}}, & k_1 \leq k_2. \end{cases}$$

In a two-stage design, (Z_1, Z_2) follows a bivariate normal distribution:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{BN} \left[\boldsymbol{\mu} = \begin{pmatrix} \theta \sqrt{\mathcal{I}_1} \\ \theta \sqrt{\mathcal{I}_2} \end{pmatrix}, \mathbf{V} = \begin{pmatrix} 1 & \sqrt{\mathcal{I}_1 / \mathcal{I}_2} \\ \sqrt{\mathcal{I}_1 / \mathcal{I}_2} & 1 \end{pmatrix} \right] \quad (13)$$

5.1. Sample sizes for the power family of one-sided group sequential tests

Suppose we wish to test $H_0: \theta = 0$ against $H_1: \theta > 0$ with type I error probability α and power $1 - \beta$ at $\theta = \delta$. Applying the power family of one-sided group sequential tests described by Jennison and Turnbull [9], the final information level is

$$\mathcal{J}_K = \frac{\{\tilde{C}_1(K, \alpha, \beta, \Delta) + \tilde{C}_2(K, \alpha, \beta, \Delta)\}^2}{\delta^2}. \quad (14)$$

The constants $\tilde{C}_1(K, \alpha, \beta, \Delta)$ and $\tilde{C}_2(K, \alpha, \beta, \Delta)$ are chosen to ensure that the type I error and power conditions be satisfied. Here, δ is a parameter that appears as a power in the formulae for their stopping boundaries. Then, the ratio of the test's maximum information level to that required for a fixed sample test is

$$\tilde{R}(K, \alpha, \beta, \Delta) = \frac{\mathcal{J}_K}{\mathcal{J}_{f,1}} = \frac{\{\tilde{C}_1(K, \alpha, \beta, \Delta) + \tilde{C}_2(K, \alpha, \beta, \Delta)\}^2}{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2},$$

where, $\mathcal{J}_{f,1}$ denotes the information level of a one-sided fixed sample test. When $K=2$, the power family one-sided group sequential test's final information level is

$$\mathcal{J}_2 = \tilde{R}(2, \alpha, \beta, \Delta) \times \mathcal{J}_{f,1}.$$

Let's assume equal allocation of disease and non-diseased samples, $n_2^D = n_2^{\bar{D}} = n_2$. It follows from equation (11) that

$$\begin{aligned} n_2 &= \mathcal{J}_2 \times (\mathbf{c}^T V \mathbf{c}), \\ V &= \begin{pmatrix} se(1-se) & 0 \\ 0 & sp(1-sp) \end{pmatrix}. \end{aligned}$$

Define $\mathbf{c} = (\lambda, \mu)^T$, then

$$n_2 = \mathcal{J}_2 \times [\lambda^2 se(1-se) + \mu^2 sp(1-sp)] = \mathcal{J}_2 \times g(se, sp).$$

Because sensitivity and specificity are not specified under the null hypothesis, the maximum value of function $g(se, sp)$ is used to obtain the sample size requirement with respect to $\theta = \theta_0$. Then, n_2 is rounded up to an integer that can be divided by 2.

In Table 5, we compare the sample size of sequential designs of testing Youden's index for $K=2$ by normal approximation method with the proposed optimal two-stage designs. We choose the power parameter $\delta = 0$, type I error probability $\alpha = 0.05$ at $\theta = 0$ and power $1 - \beta = 0.8$ at $\theta = \delta$, so $\tilde{R}(2, 0.05, 0.8, 0) = 1.073$.

When comparing the sample size of normal approximation method $n_{2,N}$ and the sample size of the optimal two-stage designs $n_{2,O}$ in Table 5, the optimal two-stage designs require no more than those of normal approximation method. $n_{2,N}$ increases in the parameter under H_0 .

It is a property of normal approximation to binomial probability distribution when the binomial parameter is close to 0 and 1. When the parameter δ decreases, $n_{2,N}$ is much larger than $n_{2,O}$. So, in the preliminary or early development phases of medical device, it is more appropriate to use exact binomial probability distribution.

6. Estimation, p-value and confidence interval of index J upon termination

Upon termination of the two-stage testing, estimation of the accuracy index J is also desirable. As is well-known, early stopping introduces bias to the maximum likelihood estimate that is otherwise unbiased in a fixed size setting [19–21], and the maximum likelihood estimate $\hat{J}_k = \hat{J}_k(S_k^D, S_k^{\bar{D}}) = \lambda S_k^D / n_k^D + \mu S_k^{\bar{D}} / n_k^{\bar{D}} - \phi$ of J is no exception. When the sample sizes $n_k^D, n_k^{\bar{D}}$ are large enough at each stage k to warrant the use of normal approximation, unbiased estimation [19] and bias-adjusted estimation [20] approach can be employed to estimate J using the formulations in the previous section. With relatively small sample sizes, exact methods are needed.

The terminal statistic $(S_k^D, S_k^{\bar{D}})$, where $k = 1$ if stopping at stage 1 and $k = 2$ if stopping at stage 2, is sufficient for making inference on J . Define

$$\Omega_1 = \{(i, j) : 0 \leq i \leq n_1^D, 0 \leq j \leq n_1^{\bar{D}}, \text{ and } \lambda \frac{i}{n_1^D} + \mu \frac{j}{n_1^{\bar{D}}} - \phi \in [-\phi, b_1]\},$$

and

$$\Omega_2 = \{(i, j) : \min(0, i_1) \leq i \leq n_2^D \text{ and } \min(0, j_1) \leq j \leq n_2^{\bar{D}} \text{ for some } (i_1, j_1) \in \bar{\Omega}_1\},$$

where $\bar{\Omega}_1$ is the complement set of Ω_1 , that is,

$$\bar{\Omega}_1 = \{(i, j) : 0 \leq i \leq n_1^D, 0 \leq j \leq n_1^{\bar{D}}, \text{ and } \lambda \frac{i}{n_1^D} + \mu \frac{j}{n_1^{\bar{D}}} - \phi > b_1\}.$$

The set Ω_k gives all possible values of $(S_k^D, S_k^{\bar{D}})$ if stopping occurs at stage k ($= 1, 2$), which in return, yields all possible values of $\hat{J}_k$.

The expectation of $\hat{J}$ can be expressed as a function of se and sp :

$$e(se, sp) = E(\hat{J}) = \sum_{(i,j) \in \Omega_1} \hat{J}_1(i, j) d(i, n_1^D, se) d(j, n_1^{\bar{D}}, sp) + \sum_{\{(i_1, j_1) \in \bar{\Omega}_1\} \cap \{(i, j) \in \Omega_2\}} \hat{J}_2(i, j) d(i_1, n_1^D, se) d(j_1, n_1^{\bar{D}}, sp) d(i - i_1, n_2^D - n_1^D, se) d(j - j_1, n_2^{\bar{D}} - n_1^{\bar{D}}, sp)$$

which can be evaluated numerically, yielding bias $h(se, sp) = \epsilon(se, sp) - J$.

A simple bias-reduced estimator can be obtained by subtracting the estimated bias, $h(\widehat{se}, \widehat{sp}) = h(\widehat{se}, \widehat{sp})$, from the estimator $\hat{J}$, where $\widehat{se}$ and $\widehat{sp}$ are the empirical estimates of se and sp upon stopping, respectively. This yields

$$J^* = \hat{J} - h(\widehat{se}, \widehat{sp}), \quad (15)$$

whose bias can be similarly evaluated numerically.

This simple idea of subtracting the estimated bias from the estimator has been shown in [22] to be quite effective in reducing bias for a two-stage phase 2 trial with a single binary outcome. Here we numerically computed the bias and root-mean-squared-error (RMSE) of the two estimators, based on selected design settings presented in Table 1 and 2. The results corresponding to the design in Table 1 with $\theta_0 = 0.3$, $\theta_1 = 0.8$, $se_1 = 0.0.9$, $\alpha = 0.05$, and $\beta = 0.2$ are presented in Table 6 and 7, for selected values of se and sp . It is noted that the bias of the likelihood estimator $\hat{J}$ can be substantial; for example, the bias is about -0.1059 when $se = 0.75$, $sp = 0.70$ (yielding $J = 0.45$). The estimator appears to underestimate the index J , with the magnitude of bias showing an increasing-then-decreasing pattern as se (or sp) increases. The bias-reduced estimator J^* considerably reduces the bias of J , for example, from -0.1059 to -0.0314 , more than 70 per cent reduction! The substantial bias reduction is incorporated with a slight increase in root-mean-squared error, a similar observation to that in [19].

Upon termination, it is also desirable to compute the p-value and confidence interval as evidence against the null hypothesis of $J = \theta_0$. Similar to group sequential testing in clinical trial literature, an ordering defining extremity in the sample space of the sufficient statistics is needed. Here we consider the ordering based on the maximum likelihood estimate $\hat{J}$ (MLE ordering); see among others, [9, 19, 23]. In addition to making the complete list of the sample space, the complication in exact calculation, as shown in the discussion of hypothesis testing and estimation of J , is the presence of se (or sp) as an nuisance parameter to the primary parameter J , that needs to be treated properly. According to the MLE ordering, a value of the sufficient statistic is more extreme if it yields more extreme value of the maximum likelihood estimate.

Now consider the two-stage procedure. Suppose the procedure terminates and yields a maximum likelihood estimate $\hat{J}$. Then the extreme sample paths are given by

$$\Lambda_1 = \left\{ (i, j) \in \Omega_1 : \hat{J}_1(i, j) > \hat{J} \right\}$$

and

$$\Lambda_2 = \{(i, j) \in \Omega_2 : \hat{J}_2(i, j) > \hat{J}\}.$$

The probability of a sample path in Λ is a function of se , $J = \theta$ and $\hat{J}$, and is given by

$$\begin{aligned} \omega(\hat{J}; \theta, se) &= \sum_{\{(i,j) \in \Lambda_1\}} d(i, n_1^D, se) d(j_1, n_1^{\overline{D}}, sp) \\ &+ \sum_{\{(i_1, j_1) \in \overline{\Omega}_1\} \cap \{(i,j) \in \Lambda_2\}} d(i_1, n_1^D, se) d(j_1, n_1^{\overline{D}}, sp) d(i \\ &- i_1, n_2^D \\ &- n_1^D, se) d(j \\ &- j_1, n_2^{\overline{D}} \\ &- n_1^{\overline{D}}, sp), \end{aligned}$$

with $sp = (\theta + \varphi - \lambda se)/\mu$.

To deal with the nuisance parameter se , define $\zeta(\theta) = \max_{se} \omega(\hat{J}; \theta, se)$, where the maximization is taken over all possible values of se , that is, $\max\{0, (\varphi + \theta - \mu)/\lambda\} \leq se \leq \min\{1, (\varphi + \theta)/\lambda\}$. Then from Bickel and Doksum [24], the p-value can be computed as $\zeta(\theta_0)$. A 100(1 - α) per-cent level confidence interval (J_l, J_u) can also be obtained by finding J_l and J_u such that $\zeta(J_u) = 1 - \zeta(J_l) = \alpha/2$.

Let us consider a simple example from the two-stage design used for Table 6 and 7. Suppose at stage 1, we have $S_1^D = 3$ and $S_1^{\overline{D}} = 3$, thus the test continues to stage 2 at the end of which we observe $S_2^D = 9$ and $S_2^{\overline{D}} = 9$, yielding $\hat{J} = 0.9$, indicating that the diagnostic biomarker is quite promising. There are no extreme cases at stage 1 since all six subjects are correctly classified. At stage 2, the more extreme cases are $(S_2^D, S_2^{\overline{D}}) = \{(9, 10), (10, 9), (10, 10)\}$. Note that $0.3 \leq se \leq 1$, the p-value is calculated to be 0.0015, resulting in rejection of the null hypothesis. A 95 per cent confidence interval is found to be [0.5301, 0.9846] whose width is relatively large due to the small sample sizes.

7. Example

We illustrate our method with an example of testing the efficiency of recent HIV-1 infection screening test, which is the probability of correctly classifying subjects by the screening test according to the true status of their condition. Many epidemiology studies conducted in different regions indicate that the prevalence of HIV-1 infection varies in age groups and demographic regions, and mobile population is at high risk of HIV-1 infection, such as fishermen [25].

A new diagnostic test for HIV-1 infection should have 90 per cent efficiency and 95 per cent sensitivity to be considered further, and if the efficiency of such test is below 75 per cent, then it is not worth studying further. Thus, we test $H_0: J = 0.75$ with adequate power at $J = 0.9$, assuming that the prevalence of HIV-1 infection in the mobile population is 20 per cent, which is thought to be at high risk for HIV-1 infection. Then $J = 0.2 \times se + (1 - 0.2) \times sp$. The significance level of the test is no more than 5 per cent and the power at 95 per cent sensitivity and 90 per cent efficiency is at least 80 per cent. When every one infected subject is tested, there are four non-infected subjects also enrolled in the study, i.e., $n^D: n^{\bar{D}} = 1:4$, then a fixed size study requires 10 HIV-1 infected subjects and 40 non-infected subjects. An optimal two-stage design needs to recruit 4 HIV-1 infected subjects and 16 non-infected subjects in stage 1. If the estimate of the efficiency exceed 0.85 in stage 1, additional 7 infected and 28 non-infected subjects are tested in stage 2. If the estimated efficiency based on the cumulative 11 HIV-1 infected and 44 non-infected subjects in stage 2 exceeds 0.84, then it is concluded that the new HIV-1 infection screening test has sufficient efficiency that warrants further evaluation. The expected sample number of this optimal two-stage design is 23.19 when the efficiency is actually 75 per cent, which is about half of the total number of subjects in a fixed size design.

8. Discussion

We proposed a sequential testing procedure for a general linear combination of the two standard measures of diagnostic accuracy, sensitivity and specificity. In the research of developing diagnostic biomarkers, it is crucial to save human medical samples if the diagnostic biomarkers are unpromising. In the earlier phases of evaluation, confining to such a one-dimension summary index can potentially save resources. The proposed sequential designs usually require larger sample sizes in order to achieve the same level of significance and power than the fixed size designs, yet the maximum expected sample sizes of the optimal two-stage designs are much less than those of the fixed size designs when indeed the diagnostic biomarker fails to meet the accuracy requirements.

Heitjan [26] modelled a phase III two-arm clinical trial with interim analysis as a decision problem. They proposed methods for constructing a group sequential test that is optimal for a given utility and prior distribution. Similar Bayesian approaches can apply to the diagnostic studies with interim analysis. An optimal group sequential test of J may be constructed for a given utility and prior distributions.

Sequential design have been considered in the field of diagnostic medicine for various purposes other than hypotheses testing. Irwig [27] described a method of sampling from a pre-assembled cohort for which results of the candidate biomarker are available. The probability of selection into the sample differs in the two strata defined by test outcome, and is chosen to minimize the variance of the estimate of sensitivity. Obuchowski and Zhou [28] considered two-stage designs in which all subjects testing positive in the first stage, as well as a random sample of test negatives, are selected for disease verification in the second stage. McNamee [29] considered both the case of sampling from a pre-assembled cohort and the case of planning a two-stage study in which the size of the cohort who received the candidate biomarker is a design parameter. Wruck [30] proposed estimators based on two-

step sequential sampling scheme in a large cohort study, in which the first step of sampling is conducted to efficiently estimate specificity, then the second step of sampling is conducted to estimate sensitivity if the candidate biomarker is determined to be of sufficient specificity. Based on the asymptotic normal approximations for the sampling distributions of $\text{logit } \widehat{se}$ and $\text{logit } \widehat{sp}$, the two-step design requires the sample sizes of a cohort study ranging from 800 to 2300. Such a large number of subjects are not available in the early research development.

Youden's index does not take into account the prevalence of disease and therefore it contains less information than other indices such as index of predictive values or efficiency. The importance of prevalence in determining operational characteristics of diagnostic biomarkers is widely recognized in practice [31]. With the incorporation of the prevalence of the disease, predictive value is clinically the most relevant value for an individual using the results of a biomarker. However, the decision of whether a new diagnostic biomarker is worthy of a larger scale study generally depends on the accuracy of the test characteristics.

The proposed two-stage procedures terminate the testing process early for futility; the process continues to the second stage only if the diagnostic biomarker demonstrates satisfactory performance at the first stage. The methods can be easily extended to early stopping for other reasons. For example, stopping at first stage for utility if the diagnostic biomarker shows effectiveness in classification. Similar technical treatments will be needed for exact calculations.

Acknowledgments

Research of A. Liu is supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH). The authors thank two anonymous referees and an associate editor for their constructive comments that resulted in an improved manuscript.

References

1. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports*. 1947; 62:1432–1449. [PubMed: 20340527]
2. Zhou, XH.; Obuchowski, NA.; Obuchowski, DM. *Statistical Methods in Diagnostic Medicine*. 1. Wiley; New York: 2002.
3. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950; 3:32–35. [PubMed: 15405679]
4. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*. 2000; 45:23–41. [PubMed: 10802332]
5. Parker RA, Davis RB. Evaluating whether a binary decision rule operates better than chance. *Biometrical Journal*. 1999; 41:25–31.
6. Biggerstaff B. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in Medicine*. 2000; 19:649–663. [PubMed: 10700737]
7. Greiner M. Two-graph receiver operating characteristics (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification costs. *Journal of Immunological Method*. 1996; 191:93–94.
8. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Chapman and Hall/CRC; New York: 2003.

9. Jennison, C.; Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. Chapman and Hall/CRC; New York: 2000.
10. Proschan, MA.; Lan, KKG.; Wittes, JT. Statistical Monitoring of Clinical Trials: A Unified Approach. Springer; New York: 2006.
11. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
12. Jung SH, Lee T, Kim KM, George SL. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine*. 2004; 23:561–569. [PubMed: 14755389]
13. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*. 1995; 51:1372–1383. [PubMed: 8589229]
14. Conaway MR, Petroni GR. Bivariate sequential designs for phase II trials. *Biometrics*. 1995; 51:656–64. [PubMed: 7662852]
15. Mazumdar M, Liu A. Group sequential design for comparative diagnostic accuracy studies. *Statistics in Medicine*. 2003; 22:727–739. [PubMed: 12587102]
16. Liu A, Wu CQ, Schisterman EF. Nonparametric sequential evaluation of diagnostic biomarkers. *Statistics in Medicine*. 2008; 27:1667–1678. [PubMed: 18241080]
17. Tang L, Emerson SS, Zhou XH. Nonparametric and semiparametric group sequential methods for comparing accuracy of diagnostic tests. *Biometrics*. 2008; 64:1137–1145. [PubMed: 18371124]
18. Shu Y, Liu A, Li Z. Sequential evaluation of a medical diagnostic test with binary outcomes. *Statistics in Medicine*. 2007; 26:4416–4427. [PubMed: 17539044]
19. Emerson EE, Fleming TR. Parameter estimation following sequential hypothesis testing. *Biometrika*. 1990; 77:875–892.
20. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*. 1997; 73:573–581.
21. Liu A, Hall WJ. Unbiased estimation following a group sequential test. *Biometrika*. 1999; 77:875–892.
22. Guo H, Liu A. A simple and efficient bias-reduced estimator of response probability following a group sequential phase II trial. *Journal of Biopharmaceutical Statistics*. 2005; 15:773781.
23. Armitage P. Numerical studies in the sequential estimation of a binomial parameter. *Biometrika*. 1958; 45:1–15.
24. Bickel, PJ.; Doksum, KA. Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day; San Francisco: 1977.
25. Entz AT, Ruffolo VP, Chinveschakitvanich V, Soskolne V, van Griensven GJ. HIV-1 prevalence, HIV-1 subtypes and risk factors among fishermen in the Gulf of Thailand and the Andaman Sea. *AIDS*. 2000; 14:1027–1034. [PubMed: 10853985]
26. Heitjan DF, Houts PS, Harvey HA. A decision-theoretic evaluation of early stopping rules. *Statistics in medicine*. 1992; 11:673–683. [PubMed: 1594809]
27. Irwig LI, Glasziou P, Berry G, Chock C, Mock P, Simpson J. Efficient study design to assess the accuracy of screening tests. *American Journal of Epidemiology*. 1994; 140:759–769. [PubMed: 7942777]
28. Obuchowski NA, Zhou XH. Prospective studies of diagnostic test accuracy when disease prevalence is low. *Biostatistics*. 2002; 3:477–492.
29. McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Statistics in Medicine*. 2002; 21:3609–3625. [PubMed: 12436459]
30. Wruck LM, Yiannoutsos CT, Hughes MD. A sequential design to estimate sensitivity and specificity of a diagnostic or screening test. *Statistics in Medicine*. 2006; 25:3458–3473. [PubMed: 16374904]
31. Fletcher, RH.; Fletcher, SW.; Wagner, EH. Clinical Epidemiology — the Essentials. Williams and Wilkins; Baltimore: 1982.

Optimal two-stage designs for testing Youden's index ($J = se + sp - 1$, $\lambda, \mu = n_k^D : n_k^D = 1:1$).

θ_0	θ_1	se_1^I	α	β	b_1	b_2	n_1/n_2^2	α^3	$1-\beta^4$	ASN_{05}^*	n_p^6
0	0.5	0.75	0.1	0.1	0.0000	0.1875	5/16	0.0847	0.9002	18.293	13
			0.05	0.2	0.1667	0.2500	6/16	0.0380	0.8083	15.877	13
0.1	0.6	0.80	0.1	0.1	0.1667	0.2857	6/14	0.0927	0.9080	16.871	12
			0.05	0.2	0.2000	0.3846	5/13	0.0383	0.8059	14.257	12
0.2	0.7	0.85	0.1	0.1	0.2000	0.4167	5/12	0.0843	0.9137	15.352	11
			0.05	0.2	0.4000	0.4286	5/14	0.0437	0.8040	13.011	12
0.3	0.8	0.9	0.1	0.1	0.4000	0.5000	5/10	0.0933	0.9107	12.616	9
			0.05	0.2	0.3333	0.6000	3/10	0.0371	0.8062	10.467	10
0.2	0.6	0.8	0.1	0.1	0.2222	0.3684	9/19	0.0960	0.9099	25.486	18
			0.05	0.2	0.2857	0.4211	7/19	0.0466	0.8050	20.702	18
0.3	0.7	0.85	0.1	0.1	0.2857	0.4706	7/17	0.0966	0.9144	22.454	16
			0.05	0.2	0.4286	0.5263	7/19	0.0384	0.8041	19.292	16
0.4	0.8	0.9	0.1	0.1	0.4000	0.5625	5/16	0.0903	0.9092	18.421	15
			0.05	0.2	0.5000	0.6429	6/14	0.0404	0.8101	16.045	14
0.5	0.9	0.95	0.1	0.1	0.6000	0.6667	5/12	0.0821	0.9019	13.416	10
			0.05	0.2	0.6000	0.7500	5/12	0.0348	0.8477	13.416	12

The pre-specified level of sensitivity under the alternative parameter space where the powers of the optimal two-stage designs and the fixed designs are evaluated.

²The sample size in both stages per group, $n_k^D = \bar{n}_k^D = n_k$, $k = 1, 2$.

³The theoretical probability of type I error achieved by the optimal two-stage designs.

⁴The theoretical power achieved by the optimal two-stage designs

⁵The maximum expected sample size for the optimal two-stage designs.

⁵The required sample size per group for fixed designs, $n^D = n^{D-} = nf$.

Optimal two-stage designs for testing efficiency ($J = \frac{1}{4}se + \frac{3}{4}sp, \lambda; \mu = n_k^D : n_k^{\overline{D}} = 1:3$).

θ_0	θ_1	se_1	α	β	b_1	b_2	$(n_1^D, n_1^D)/(n_2^D, n_2^D)$	α^*	$1-\beta$	ASN_0^*	n_f
0.5	0.8	0.9	0.1	0.1	0.5000	0.7000	(2,6)/(6,18)	0.0505	0.9129	13.647	(5,15)
			0.05	0.2	0.5000	0.6429	(1,3)/(7,21)	0.0286	0.8019	10.773	(4,12)
0.6	0.9	0.9	0.1	0.1	0.6250	0.7500	(2,6)/(4,12)	0.0605	0.9174	10.486	(4,12)
			0.05	0.2	0.7500	0.7500	(2,6)/(5,15)	0.0275	0.8014	9.277	(5,15)
0.5	0.75	0.8	0.1	0.1	0.5625	0.5938	(4,12)/(8,24)	0.0797	0.9054	19.621	(7,21)
			0.05	0.2	0.5834	0.6250	(3,9)/(8,24)	0.0380	0.8090	15.840	(7,21)
0.6	0.85	0.9	0.1	0.1	0.6667	0.7084	(3,9)/(6,18)	0.0818	0.9109	14.697	(6,18)
			0.05	0.2	0.6250	0.7500	(2,6)/(6,18)	0.0341	0.8091	12.971	(6,18)
0.6	0.8	0.9	0.1	0.1	0.7000	0.6731	(5,15)/(13,39)	0.0738	0.9051	24.019	(9,27)
			0.05	0.2	0.6667	0.7045	(3,9)/(11,33)	0.0419	0.8273	19.192	(9,27)
0.7	0.9	0.95	0.1	0.1	0.7500	0.7813	(4,12)/(8,24)	0.0868	0.9173	19.934	(8,24)
			0.05	0.2	0.7500	0.8571	(3,9)/(7,21)	0.0404	0.8114	16.045	(7,21)

Table 3

Simulation study of optimal two-stage designs of for testing Youden's Index in Table 1.

θ_0	θ_1	θ_{se}	α	α^*I	α_s^2	ASN_{03}^*	ASN_s^4
0	0.5	0.75	0.1	0.0847	0.0852	18.293	18.334
			0.05	0.0380	0.0385	15.877	15.878
0.1	0.6	0.80	0.1	0.0927	0.0945	16.871	16.895
			0.05	0.0383	0.0390	14.257	14.282
0.2	0.7	0.85	0.1	0.0843	0.0855	15.352	15.372
			0.05	0.0437	0.0444	13.011	13.049
0.3	0.8	0.9	0.1	0.0933	0.0943	12.616	12.642
			0.05	0.0371	0.0377	10.467	10.487
0.2	0.6	0.8	0.1	0.0960	0.0971	25.486	25.527
			0.05	0.0466	0.0466	20.702	20.722
0.3	0.7	0.85	0.1	0.0966	0.0995	22.454	22.490
			0.05	0.0384	0.0359	19.292	19.355
0.4	0.8	0.9	0.1	0.0903	0.0911	18.421	18.448
			0.05	0.0404	0.0409	16.045	16.055
0.5	0.9	0.95	0.1	0.0821	0.0828	13.416	13.419
			0.05	0.0348	0.0355	13.416	13.435

^{1,3}The theoretical values achieved by the optimal two-stage designs.

^{2,4}The simulated values based on 100,000 iterations.

Table 4

Simulation study of optimal two-stage designs for testing efficiency in Table 2.

θ_0	θ	$1-\epsilon$	α	α^*	α_c	ASN_0^*	ASN_s
0.5	0.8	0.9	0.1	0.0505	0.0300	13.647	13.848
			0.05	0.0286	0.0293	10.773	11.541
0.6	0.9	0.9	0.1	0.0605	0.0622	10.486	10.538
			0.05	0.0275	0.0281	9.277	9.298
0.5	0.75	0.8	0.1	0.0797	0.0811	19.621	19.667
			0.05	0.0380	0.0387	15.840	15.910
0.6	0.85	0.9	0.1	0.0818	0.0747	14.697	14.727
			0.05	0.0341	0.0353	12.971	13.086
0.6	0.8	0.9	0.1	0.0738	0.0623	24.019	25.828
			0.05	0.0419	0.0364	19.192	19.288
0.7	0.9	0.95	0.1	0.0868	0.0881	19.934	19.978
			0.05	0.0404	0.0410	16.045	16.073

Table 5

Sample sizes of one-sided group sequential tests on Youden's index with equal group size. $n_k = n_k^D = n_k^{\overline{D}}$, $k = 1, 2$ and $n_1 : n_2 = 1 : 2$. Type I error probability $\alpha = 0.05$ at $\theta = 0$ and power $1 - \beta = 0.8$ at $\theta = \delta$. The power parameter is 0.

δ	J_0^1	J_1^2	$n_{2,N}^3$	$n_{2,O}^4$
0.5	0	0.5	16	16
	0.1	0.6	16	13
	0.2	0.7	16	14
	0.3	0.8	18	10
0.4	0.2	0.6	26	19
	0.3	0.7	28	19
	0.4	0.8	30	14
	0.5	0.9	32	12

^{1,2}The value of J under H_0 and H_1 , respectively.

³The sample size per group in two-stage designs for normal approximation approach.

⁴The sample size per group in the optimal two-stage designs using exact binomial distribution.

Table 6Bias of maximum likelihood estimator ($\hat{\mathcal{J}}$) and bias-reduced estimator ($\hat{\mathcal{J}}^*$).

ds	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
$\hat{\mathcal{J}}$										
0.55	-0.0713	-0.0793	-0.0863	-0.0922	-0.0966	-0.0993	-0.0998	-0.0980	-0.0934	-0.0875
0.60	-0.0793	-0.0871	-0.0937	-0.0988	-0.1021	-0.1032	-0.1020	-0.0980	-0.0910	-0.0830
0.65	-0.0863	-0.0937	-0.0995	-0.1034	-0.1052	-0.1046	-0.1012	-0.0949	-0.0854	-0.0753
0.70	-0.0922	-0.0988	-0.1034	-0.1059	-0.1059	-0.1032	-0.0976	-0.0889	-0.0770	-0.0651
0.75	-0.0966	-0.1021	-0.1052	-0.1059	-0.1038	-0.0989	-0.0910	-0.0801	-0.0662	-0.0529
0.80	-0.0993	-0.1032	-0.1046	-0.1032	-0.0989	-0.0918	-0.0817	-0.0689	-0.0536	-0.0396
0.85	-0.0998	-0.1020	-0.1012	-0.0976	-0.0910	-0.0817	-0.0699	-0.0558	-0.0399	-0.0263
0.90	-0.0980	-0.0980	-0.0949	-0.0889	-0.0801	-0.0689	-0.0558	-0.0413	-0.0262	-0.0143
0.95	-0.0934	-0.0910	-0.0854	-0.0770	-0.0662	-0.0536	-0.0399	-0.0262	-0.0135	-0.0050
0.99	-0.0875	-0.0830	-0.0753	-0.0651	-0.0529	-0.0396	-0.0263	-0.0143	-0.0050	-0.0007
$\hat{\mathcal{J}}^*$										
0.55	-0.0186	-0.0225	-0.0258	-0.0285	-0.0303	-0.0312	-0.0310	-0.0296	-0.0271	-0.0241
0.60	-0.0225	-0.0261	-0.0290	-0.0310	-0.0320	-0.0319	-0.0306	-0.0281	-0.0245	-0.0207
0.65	-0.0258	-0.0290	-0.0312	-0.0323	-0.0323	-0.0311	-0.0288	-0.0252	-0.0206	-0.0161
0.70	-0.0285	-0.0310	-0.0323	-0.0325	-0.0314	-0.0291	-0.0257	-0.0212	-0.0158	-0.0110
0.75	-0.0303	-0.0320	-0.0323	-0.0314	-0.0292	-0.0258	-0.0215	-0.0163	-0.0106	-0.0057
0.80	-0.0312	-0.0319	-0.0311	-0.0291	-0.0258	-0.0216	-0.0166	-0.0111	-0.0054	-0.0009
0.85	-0.0310	-0.0306	-0.0288	-0.0257	-0.0215	-0.0166	-0.0112	-0.0059	-0.0008	0.0027
0.90	-0.0296	-0.0281	-0.0252	-0.0212	-0.0163	-0.0111	-0.0059	-0.0011	0.0026	0.0046
0.95	-0.0271	-0.0245	-0.0206	-0.0158	-0.0106	-0.0054	-0.0008	0.0026	0.0044	0.0043
0.99	-0.0241	-0.0207	-0.0161	-0.0110	-0.0057	-0.0009	0.0027	0.0046	0.0043	0.0020

Table 7

Root-mean-squared-error (RMSE) of maximum likelihood estimator ($\hat{\mathcal{J}}$) and bias-reduced estimator ($\hat{\mathcal{J}}^*$).

ds	se									
	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99
$\hat{\mathcal{J}}$										
0.55	0.3325	0.3266	0.3203	0.3134	0.3059	0.2973	0.2872	0.2748	0.2590	0.2429
0.60	0.3266	0.3220	0.3171	0.3117	0.3057	0.2984	0.2893	0.2775	0.2618	0.2451
0.65	0.3203	0.3171	0.3137	0.3097	0.3049	0.2985	0.2898	0.2779	0.2613	0.2432
0.70	0.3134	0.3117	0.3097	0.3070	0.3029	0.2969	0.2880	0.2752	0.2567	0.2363
0.75	0.3059	0.3057	0.3049	0.3029	0.2992	0.2929	0.2831	0.2684	0.2472	0.2237
0.80	0.2973	0.2984	0.2985	0.2969	0.2929	0.2856	0.2741	0.2568	0.2320	0.2046
0.85	0.2872	0.2893	0.2898	0.2880	0.2831	0.2741	0.2599	0.2394	0.2102	0.1782
0.90	0.2748	0.2775	0.2779	0.2752	0.2684	0.2568	0.2394	0.2147	0.1806	0.1435
0.95	0.2590	0.2618	0.2613	0.2567	0.2472	0.2320	0.2102	0.1806	0.1411	0.0987
0.99	0.2429	0.2451	0.2432	0.2363	0.2237	0.2046	0.1782	0.1435	0.0987	0.0504
$\hat{\mathcal{J}}^*$										
0.55	0.3585	0.3504	0.3411	0.3308	0.3192	0.3063	0.2917	0.2749	0.2553	0.2369
0.60	0.3504	0.3428	0.3343	0.3248	0.3141	0.3019	0.2879	0.2716	0.2522	0.2336
0.65	0.3411	0.3343	0.3266	0.3179	0.3079	0.2963	0.2826	0.2661	0.2462	0.2269
0.70	0.3308	0.3248	0.3179	0.3099	0.3003	0.2889	0.2750	0.2579	0.2369	0.2162
0.75	0.3192	0.3141	0.3079	0.3003	0.2909	0.2792	0.2646	0.2464	0.2236	0.2010
0.80	0.3063	0.3019	0.2963	0.2889	0.2792	0.2668	0.2510	0.2310	0.2059	0.1808
0.85	0.2917	0.2879	0.2826	0.2750	0.2646	0.2510	0.2334	0.2112	0.1832	0.1552
0.90	0.2749	0.2716	0.2661	0.2579	0.2464	0.2310	0.2112	0.1862	0.1548	0.1235
0.95	0.2553	0.2522	0.2462	0.2369	0.2236	0.2059	0.1832	0.1548	0.1197	0.0845
0.99	0.2369	0.2336	0.2269	0.2162	0.2010	0.1808	0.1552	0.1235	0.0845	0.0436