Published in final edited form as:

Stat Med. 2012 December 10; 31(28): 3638–3648. doi:10.1002/sim.5391.

Homogeneity tests of clustered diagnostic markers with applications to the BioCycle Study

Liansheng Larry Tang^{a,†,*}, Aiyi Liu^b, Enrique F. Schisterman^b, Xiao-Hua Zhou^c, and Catherine Chun-ling Liu^d

^aDepartment of Statistics, George Mason University, Fairfax, VA 22030, U.S.A

^bDivision of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, Rockville, MD 20852, U.S.A

^cDepartment of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A

^dDepartment of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

Abstract

Diagnostic trials often require the use of a homogeneity test among several markers. Such a test may be necessary to determine the power both during the design phase and in the initial analysis stage. However, no formal method is available for the power and sample size calculation when the number of markers is greater than two and marker measurements are clustered in subjects. This article presents two procedures for testing the accuracy among clustered diagnostic markers. The first procedure is a test of homogeneity among continuous markers based on a global null hypothesis of the same accuracy. The result under the alternative provides the explicit distribution for the power and sample size calculation. The second procedure is a simultaneous pairwise comparison test based on weighted areas under the receiver operating characteristic curves. This test is particularly useful if a global difference among markers is found by the homogeneity test. We apply our procedures to the BioCycle Study designed to assess and compare the accuracy of hormone and oxidative stress markers in distinguishing women with ovulatory menstrual cycles from those without.

Keywords

ROC curve; biomarker; homogeneity test; sample size

1. Introduction

Diagnostic trials are often carried out to test and compare the diagnostic accuracy of markers. Elegant methods have been proposed to combine markers after the trials are finished (see, for example, [1, 2]). However, these methods may not be applicable when the

Copyright © 2012 John Wiley & Sons, Ltd.

*Correspondence to: Liansheng Larry Tang, Department of Statistics, George Mason University, Fairfax, VA 22030, U.S.A.

†ltang1@gmu.edu

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institutes of Health, or the Department of Veterans Affairs.

investigators try to determine the power during the design phase and in the initial analysis stage. A homogeneity test offers a better alternative. When more than two markers are measured on the same set of diseased and non-diseased subjects, how to appropriately obtain the test statistic and its variance becomes challenging.

The BioCycle Study [3, 4] is a longitudinal study conducted at the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development of the National Institutes of Health. One of the primary objectives of the study is to assess and compare the accuracy of endogenous hormones (i.e., estrogen and progesterone) and oxidative stress markers during the menstrual cycle in distinguishing women with ovulatory menstrual cycles from those without. This helps further the understanding of the implications of oxidative stress in the risk of human infertility and the mechanisms by which oxidative stress may be associated with female infertility. The study enrolled healthy premenopausal women and measured both oxidative stress markers and endogenous hormones in serum on multiple days [3].

In this study, we initially classified menstrual cycles as anovulatory if the peak progesterone concentration across the cycle was <5 ng/mL. To minimize misclassification, we employed a more conservative definition of anovulation in which cycles with progesterone concentrations ≤ 5 ng/mL and an observed serum luteinizing hormone (LH) peak in the mid or late luteal phase visit were considered ovulatory. On the basis of this algorithm, 35 out of 259 women had at least one anovulatory cycle, and the rest of the women were ovulatory. The investigators were interested in comparing three hormones [estradiol (E2), follicle stimulating hormone (FSH), and LH and an oxidative stress marker F2isoprostanes (F2iso) on the visits around ovulation (corresponding to the late follicular phase, LH/FSH surge, and predicted ovulation) to see whether these markers have different diagnostic accuracy and, if so, which of these markers are better. Figure 1 shows the histograms of four markers' levels at the first visit around ovulation. We can see that it is difficult to make a reasonable parametric assumption about the distributions of marker measurements in our study. It is also impossible to transform the marker observations to follow some known distributions because part (a) in Figure 1 clearly indicates that observations from the anovulatory group and the ovulatory group tend to skew differently. Therefore, a nonparametric or semiparametric test that does not require any distribution assumptions would be more appropriate than parametric methods.

The primary scientific questions to be answered in the BioCycle Study include the following: (1) testing homogeneity among markers and (2) power and sample size calculation for this study. However, to the best of our knowledge, no method is available to address these two questions. Although we may extend semiparametric kernel ROC methods proposed in [5] to compare multiple markers by obtaining test statistics under the null hypothesis using some resampling methods, it is difficult to carry out the resampling under the alternative. It is also infeasible to conduct resampling at the design stage when no data are available. In this paper, we will introduce a formal homogeneity test for testing diagnostic accuracy among markers. The proposed procedure tests a global hypothesis. The χ^2 distribution derived under the alternative for the test statistic is useful for power analysis based on hypothesized differences among markers. When the number of markers is only two, the proposed global test reduces to the commonly used test by [6]. In addition, we will

also propose a simultaneous pairwise comparison procedure based on generalized ROC summary statistics. This procedure is applicable if the test of homogeneity rejects the null hypothesis and a subsequent analysis is needed to search for the set of most accurate markers. The contributions of the proposed tests are not limited to the aforementioned example. The tests are also applicable in comparing multiple medical imaging modalities which are of wide interest in radiological studies (see, for example, [7, 8]).

The rest of the article is organized as follows. In Section 2, we propose a global test based on the hypothesis of homogeneous accuracy and discuss its theoretical properties. In Section 3, we introduce a simultaneous pairwise comparison test if significance is found by the homogeneity test. We also investigate the theoretical property of the procedure. We apply the proposed procedures to the BioCycle Study data in Section 4. Section 5 reports simulation results to illustrate the small sample performance of the proposed method. We defer the proof of the asymptotic distributions to the Appendix.

2. A test of homogeneity

In this section, we introduce a test of homogeneity based on ROC summary statistics for clustered observations. We start with a brief introduction to ROC summary statistics. We then use these summary statistics to construct the test.

Let $X_{\ell ip}$ denote test result of marker ℓ on day p in the diseased subject i , where $\ell = 1, \dots, L$, $p = 0, 1, \dots, m_{\ell i}$, and $i = 1, \dots, I$. Let $Y_{\ell jq}$ denote test result of ℓ th marker on day q in the non-diseased subject j , where $\ell = 1, \dots, L$, $q = 0, 1, \dots, n_{\ell j}$, and $j = 1, \dots, J$. Here, the total number of subjects is $N = I + J$. Define the joint survival function $(X_{\ell 1p1}, X_{\ell 2p2}) \sim S_{D, \ell, 1, \ell_2}(x_1, x_2)$, $p_1 = 1, \dots, m_{\ell 1i}$, $p_2 = 1, \dots, m_{\ell 2i}$, for the diseased subjects with marginal survival functions $X_{\ell ip} \sim S_{D, \ell}(x)$. Similarly, define $(Y_{\ell 1q1}, Y_{\ell 2q2}) \sim S_{D, \ell, 1, \ell_2}(y_1, y_2)$, $q_1 = 1, \dots, n_{\ell 1j}$, $q_2 = 1, \dots, n_{\ell 2j}$ for the non-diseased subjects with marginal survival functions $Y_{\ell jq} \sim S_{D, \ell}(y)$. Without loss of generality, we assume that measurements tend to be larger for the diseased than for the non-diseased. Let u be the false positive rate (FPR), which is also 1-specificity. The

ROC curve for the ℓ th marker is $\text{ROC}_{\ell}(u) = S_{D, \ell}\{S_{D, \ell}^{-1}(u)\}$, where $u \in [0, 1]$. The resulting ℓ th weighted area under curve (wAUC) is

$$\Omega_{\ell} = \int_0^1 [S_{D, \ell}\{S_{D, \ell}^{-1}(u)\}] dW(u), \quad (1)$$

with a probability measure $W(u)$ defined on u , for $u \in (0, 1)$. Included in this class of accuracy measures are AUC (when $W(u) = u$ for $0 < u < 1$), partial AUC (pAUC) between FPRs u_1 and u_2 (when $W(u) = (u - u_1)/(u_2 - u_1)$ for $0 < u_1 < u < u_2 < 1$), and the sensitivity at a given level of FPR u_0 (when $W(u)$ is a point mass at u_0). The nonparametric wAUC

estimator is given by $\hat{\Omega}_{\ell} = \int_0^1 [\hat{S}_{D, \ell}\{\hat{S}_{D, \ell}^{-1}(u)\}] dW(u)$, where $\hat{S}_{D, \ell}$ and $\hat{S}_{D, \ell}^{-1}$ are empirical functions of $S_{D, \ell}$ and $S_{D, \ell}$ defined by

$$\hat{S}_{D,\ell}(x) = 1 - \frac{1}{\sum_{i=1}^I m_{\ell i}} \sum_{i=1}^I \sum_{p=1}^{m_{\ell i}} I(X_{\ell ip} \leq x),$$

and

$$\hat{S}_{\bar{D},\ell}(x) = 1 - \frac{1}{\sum_{j=1}^J n_{\ell j}} \sum_{j=1}^J \sum_{q=1}^{n_{\ell j}} I(Y_{\ell jq} \leq x).$$

Denote $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_L)$. By substituting $\hat{S}_{D,\ell}$ and $\hat{S}_{\bar{D},\ell}$ in Equation (1), the nonparametric estimator of Ω is given by $\hat{\Omega} = (\hat{\Omega}_1, \hat{\Omega}_2, \dots, \hat{\Omega}_L)$.

2.1. Comparing multiple diagnostic tests

We will consider the following null hypothesis of homogeneity among markers:

$$H_0: \Omega_1 = \Omega_2 = \dots = \Omega_L.$$

This hypothesis is an important initial step in comparing multiple markers [9]. Rejecting this null hypothesis is indicative of significant difference among some markers. Subsequent analysis such as pairwise comparisons can then be conducted to search for the optimal set of markers which have the best diagnostic accuracy. To construct a test statistic for the null hypothesis, we define a new vector as

$$\Omega_c = (\Omega_{c,1}, \Omega_{c,2}, \dots, \Omega_{c,L-1}),$$

where $\Omega_{c,\ell} = \Omega_\ell - \Omega_L$, for $\ell < L$. In fact, any of the markers can serve as marker L to construct the test statistic. This is due to the fact that the hypothesis of homogeneity is equivalent to testing $H_0: \Omega_c = \mathbf{0}$ vs. $H_a: \Omega_c \neq \mathbf{0}$. A consistent estimator of Ω_c is given by $\hat{\Omega}_c = (\hat{\Omega}_1 - \hat{\Omega}_L, \hat{\Omega}_2 - \hat{\Omega}_L, \dots, \hat{\Omega}_{L-1} - \hat{\Omega}_L)'$. The relationship between $\hat{\Omega}$ and $\hat{\Omega}_c$ can be expressed by $\hat{\Omega}_c = \mathbf{M}\hat{\Omega}$, where $\mathbf{M} = (\mathbf{I}_{L-1}, -\mathbf{1}_{L-1})$, with an identity matrix $\mathbf{I}_{L-1}$ and a vector of one's $\mathbf{1}_{L-1} = (1, 1, \dots, 1)'$.

Denote $m_\ell = \sum_{i=1}^I m_{\ell i}$ and $n_\ell = \sum_{j=1}^J n_{\ell j}$. Assume that $I/J \rightarrow \lambda < \infty$, as $I, J \rightarrow \infty$. Let

$$r_\ell(u) = S'_{D,\ell} \{S_{\bar{D},\ell}^{-1}(u)\} / S'_{\bar{D},\ell} \{S_{D,\ell}^{-1}(u)\}, \quad \text{for } \ell = 1, \dots, L,$$

where $S'_{D,\ell}$ and $S'_{\bar{D},\ell}$ are the first derivatives of $S_{D,\ell}$ and $S_{\bar{D},\ell}$, respectively.

Define

$$W_{\ell,i} = m_{\ell}^{-1/2} \sum_{p=1}^{m_{\ell i}} \int [I\{X_{\ell ip} > \hat{G}_{\ell}^{-1}(u)\} - \hat{S}_{D,\ell}\{\hat{G}_{\ell}^{-1}(u)\}] dW(u),$$

and

$$V_{\ell,j} = n_{\ell}^{-1/2} \sum_{q=1}^{n_{\ell j}} \int \hat{r}_{\ell}(u) [I\{X_{\ell jq} > \hat{G}_{\ell}^{-1}(u) - u\}] dW(u),$$

and $\hat{r}_{\ell}(u)$ is the kernel density estimate of $r_{\ell}(u)$. In [10], the authors used the Epanechnikov kernel function $E(x) = \frac{3}{4}(1-x^2)I(|x| \leq 1)$ with the bandwidth of $4/\max(\min(I, J)^{4/5}, 50)$, where I and J are sample sizes for the diseased group and the non-diseased group, respectively. The same setting was used in our simulation studies and the example. Other kernel functions such as the Gaussian kernel may also be used. A detailed discussion on kernel methods can be found in [11]. Denote

$$\tilde{\mathbf{W}}_i = \begin{pmatrix} W_{1,i} - W_{L,i} \\ W_{2,i} - W_{L,i} \\ \vdots \\ W_{L-1,i} - W_{L,i} \end{pmatrix}, \text{ and } \tilde{\mathbf{V}}_j = \begin{pmatrix} V_{1,j} - V_{L,j} \\ V_{2,j} - V_{L,j} \\ \vdots \\ V_{L-1,j} - V_{L,j} \end{pmatrix}.$$

In the Appendix, we show that $\hat{\boldsymbol{\Omega}}_c$ asymptotically follows a multivariate normal distribution given by

$$N_L(\boldsymbol{\Omega}_c, \sum_c = E(\tilde{\mathbf{W}}_i \tilde{\mathbf{W}}_i')/I + E(\tilde{\mathbf{V}}_j \tilde{\mathbf{V}}_j')/J). \quad (2)$$

A nonnegative definite covariance estimator of $\boldsymbol{\Sigma}_c$ is

$$\hat{\sum}_c = I^{-2} \sum_{i=1}^I \tilde{\mathbf{W}}_i \tilde{\mathbf{W}}_i' + J^{-2} \sum_{j=1}^J \tilde{\mathbf{V}}_j \tilde{\mathbf{V}}_j'. \quad (3)$$

The asymptotic distribution of the test statistic $T_c := \hat{\boldsymbol{\Omega}}_c^{-1} \hat{\sum}_c^{-1}$ is provided in the following theorem and is proved in the Appendix.

Theorem 1—With mild regularity conditions, T_c converges in distribution to a χ^2 distribution with d.f. L minus 1 under H_0 , and T_c converges in distribution to a noncentral chi-square distribution $\chi_{L-1}^2(\phi)$ distribution with the non-centrality parameter,

$$\phi = \Omega_c' \sum_c^{-1} \Omega_c,$$

under the alternative hypothesis, as $I, J \rightarrow \infty$.

For $L = 2$, Σ_c reduces to one element, $\text{var}(\hat{\Omega}_1 - \hat{\Omega}_2)$. The asymptotic distribution of $T_c = (\hat{\Omega}_1 - \hat{\Omega}_2)^2 / \text{var}(\hat{\Omega}_1 - \hat{\Omega}_2)$ reduces to a $\chi_1^2(\phi)$ distribution with the non-centrality $\phi = c^2 / \text{var}(\hat{\Omega}_1 - \hat{\Omega}_2)$ under the alternative H_a : $\Omega_1 - \Omega_2 = c$. It implies that $(\hat{\Omega}_1 - \hat{\Omega}_2) / \sqrt{\text{var}(\hat{\Omega}_1 - \hat{\Omega}_2)}$ has an approximately normal distribution with mean $\sqrt{\phi}$ and variance 1 under H_a and has approximately a standard normal distribution under H_0 . This asymptotic normality has been studied by [12] and [13]. Their expression of $\text{var}(\hat{\Omega}_1 - \hat{\Omega}_2)$ is the same as ours.

When different subjects are measured with different markers, $\hat{\Omega}_\ell$'s are independent estimators, and we can derive an explicit form for the variance-covariance matrix of $\hat{\Omega}_c$. Assume that $I/m_\ell \rightarrow \alpha_\ell$, and $I/n_\ell \rightarrow \beta_\ell$, as $I, J \rightarrow \infty$. Here, α_ℓ and β_ℓ are finite numbers. We also assume that $\sum_{i=1}^I m_{\ell_1 i} m_{\ell_2 i} / I^2 \rightarrow \eta_{\ell_1, \ell_2}^X$, and $\sum_{j=1}^J n_{\ell_1 j} n_{\ell_2 j} / I^2 \rightarrow \eta_{\ell_1, \ell_2}^Y$, as $I, J \rightarrow \infty$. It follows that

$$\Sigma_c = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_{L-1} \end{pmatrix} + v_L \mathbf{1}\mathbf{1}',$$

where $\mathbf{1} = (1, 1, \dots, 1)'$, and $v_\ell = \text{var}(\hat{\Omega}_\ell)$, whose explicit form is given in Theorem 2 by letting $h(\cdot) = \Omega_\ell$,

$$\begin{aligned} v_\ell = & I^{-1} \sum_\ell \alpha_\ell^2 \eta_{\ell, \ell}^x \left(\int_0^1 \int_0^1 [S_{D, \ell, \ell} \{S_{\bar{D}, \ell}^{-1}(s), S_{\bar{D}, \ell}^{-1}(t)\} - S_{D, \ell} \{S_{\bar{D}, \ell_1}(s)\} S_{D, \ell} \{S_{\bar{D}, \ell}^{-1}(t)\}] dW(s) dW(t) \right) \\ & + I^{-1} \sum_\ell \beta_\ell^2 \eta_{\ell, \ell}^y \left(\int_0^1 \int_0^1 r_\ell(s) r_\ell(t) [S_{\bar{D}, \ell, \ell} \{S_{\bar{D}, \ell}^{-1}(s), S_{\bar{D}, \ell}^{-1}(t)\} - st] dW(s) dW(t) \right). \end{aligned} \quad (4)$$

Given the special form of the variance-covariance matrix Σ_c , the inverse matrix $\sum_c^{-1}$ is given by an explicit expression using results in matrix operations. Theorem 1 then gives a simplified distribution result of T_c , which is summarized in the following corollary.

Corollary 1—With mild regularity conditions, as $I, J \rightarrow \infty$, T_c converges in distribution to a χ^2 distribution, with $L - 1$ degree of freedom under H_0 , and converges in distribution to a noncentral chi-square distribution $\chi_{L-1}^2(\phi)$ with

$$\phi = \sum_{\ell=1}^{L-1} v_{\ell}^{-1} (\Omega_{\ell} - \Omega_L)^2 - \left\{ \sum_{\ell=1}^{L-1} v_{\ell}^{-1} (\Omega_{\ell} - \Omega_L) \right\}^2 / \sum_{\ell=1}^L v_{\ell}^{-1}$$

under H_a .

2.2. Power and sample size

The explicit forms of χ^2 distributions under H_a in either Theorem 1 or Corollary 1 are useful for power analysis. The power under H_a is given by

$$P(X^2 > \chi_{L-1, \alpha}^2),$$

where $X^2 \sim \chi_{L-1}^2(\phi)$, and $\chi_{L-1, \alpha}^2$ is the upper α critical value of a chi-square distribution

with $L - 1$ degrees of freedom. The non-centrality ϕ can be estimated by $\hat{\phi} = \hat{\Omega}_c' \sum_c^{-1} \hat{\Omega}_c$, where the hypothesized value of Ω_c is specified under the alternative. Asymptotically, $\hat{\phi}$ converges to its true value.

For a given power $1 - \beta$, we have the following relationship

$$\chi_{L-1, \beta}^2(\phi) = \chi_{L-1, \alpha}^2, \quad (5)$$

where $\chi_{L-1, \beta}^2(\phi)$ is the upper β critical value of a noncentral chi-square distribution with $L - 1$ degrees of freedom. By fixing the power $1 - \beta$ and type I error rate α , the non-centrality parameter ϕ can be solved numerically to satisfy the relationship in (5). Let the solution for ϕ be $\phi_{\alpha, 1-\beta}$. By combining the expression of ϕ in Theorem 1 and the expression of Σ_c in (2), it follows that

$$\phi_{\alpha, 1-\beta} = \Omega_c' [E(\tilde{W}_i \tilde{W}_i') / I + E(\tilde{V}_j \tilde{V}_j') / J]^{-1} \Omega_c. \quad (6)$$

The minimal relevant difference, Ω_c , among ROC summary measures can be specified by investigators. Using either pilot studies or specific distribution assumptions, we can estimate

$E(\tilde{W}_i \tilde{W}_i')$ and $E(\tilde{V}_j \tilde{V}_j')$. By specifying a ratio between I and J , we can then solve (6) for the required sample sizes for the diseased and non-diseased patients.

3. Post hoc pairwise comparisons based on χ^2 -statistics

After the test finds a significant difference among markers, we can use a post hoc pairwise comparison to identify the most accurate markers by using differences between the weighted AUCs, $\ell_k = \Omega_{\ell} - \Omega_k$, for $\ell \neq k$. The estimator $\hat{\ell}_k = \hat{\Omega}_{\ell} - \hat{\Omega}_k$, for $\ell \neq k$ can be obtained from the data. A general setting for $\hat{\Omega}_{\ell} - \hat{\Omega}_k$ is when $m_{\ell i}, n_{\ell j} \geq 1$. In this paper, the multivariate

normality of $\hat{\Omega}_c$ in (8) of the Appendix allows derivation of the explicit expression of the asymptotic variance of $\hat{\ell}_k$.

Theorem 2—With mild regularity conditions,

$$\sqrt{I}(\hat{\Delta}_{\ell k} - \hat{\Delta}_{\ell k}) \xrightarrow{d} \mathcal{N}(0, v_X + \lambda v_Y), \quad \text{as } I, J \rightarrow \infty,$$

where

$$v_X = \sum_{\ell=1}^2 \alpha_{\ell}^2 \eta_{\ell, \ell}^x \left(\int_0^1 \int_0^1 S_{D, \ell} \{S_{\bar{D}, \ell}^{-1}(s \wedge t)\} dW(s) dW(t) - \left[\int_0^1 S_{D, \ell} \{S_{\bar{D}, \ell}^{-1}(s)\} dW(s) \right]^2 \right) \\ - 2\alpha_1 \alpha_2 \eta_{1,2}^x \int_0^1 \int_0^1 \left[S_{D,1,2} \{S_{\bar{D},1}^{-1}(s), S_{\bar{D},2}^{-1}(t)\} - S_{D,1} \{S_{\bar{D},1}^{-1}(s)\}, S_{D,2} \{S_{\bar{D},2}^{-1}(t)\} \right] dW(s) dW(t),$$

and

$$v_Y = \sum_{\ell=1}^2 \beta_{\ell}^2 \eta_{\ell, \ell}^y \left[\int_0^1 \int_0^1 r_{\ell}(s) r_{\ell}(t) (s \wedge t) dW(s) dW(t) - \left\{ \int_0^1 r_{\ell}(s) dW(s) \right\}^2 \right] \\ - 2\beta_1 \beta_2 \eta_{1,2}^y \int_0^1 \int_0^1 r_1(s) r_2(t) [S_{\bar{D},1,2} \{S_{\bar{D},1}^{-1}(s), S_{\bar{D},2}^{-1}(t)\} - st] dW(s) dW(t),$$

where

$$r_{\ell}(u) = S_{D, \ell} \{S_{\bar{D}, \ell}^{-1}(u)\} / S_{\bar{D}, \ell} \{S_{\bar{D}, \ell}^{-1}(u)\}, \quad \text{for } \ell=1, \dots, L.$$

Theorem 2 is a direct result obtained by combining the multivariate normal distribution of $\hat{\Omega}$ and the Cramer–Wold device [14, Theorem 1.5.2]. Special details are omitted here.

For the purpose of pairwise comparisons after the simultaneous test, $L-1$ -statistics can be defined to be the difference between markers. That is, $\hat{\Omega}_{1L} = \hat{\Omega}_1 - \hat{\Omega}_L$, $\hat{\Omega}_{2L} = \hat{\Omega}_2 - \hat{\Omega}_L$, ..., $\hat{\Omega}_{L-1,L} = \hat{\Omega}_{L-1} - \hat{\Omega}_L$. By directly using results in Theorem 2, p -values can be obtained for these pairwise comparisons. By comparing these p -values to a certain threshold from multiple comparison procedures such as the Bonferroni test or false discovery rate method, markers which have better accuracy can be identified.

The pairwise comparison after the proposed homogeneity test in diagnostic trials is analogous to post hoc tests after the analysis of variance test in clinical trials. In our setting, we use an overall significance test to monitor the diagnostic trial, for example, to determine early stopping of the trial. After the trial is terminated on the basis of the global test, methods such as a step-down approach are then employed to identify individual biomarkers.

4. The BioCycle Study revisited

In the aforementioned BioCycle Study, 35 women were anovulatory, and 224 women were ovulatory. The markers were measured at three visits for all patients. The investigators were interested in comparing accuracy between markers of E2, FSH, LH, and F2iso in distinguishing between ovulatory and anovulatory menstrual cycles using ROC summary measures. The empirical ROC curves of these markers are shown in Figure 2(a). The ROC derivative function $r(u)$ using kernel density smoothing estimation for each marker is illustrated in Figure 2(b). We first compared these markers on the basis of their AUCs. We conducted a homogeneity test of all these markers using results from Theorem 1 with $\alpha = 0.05$. We estimated the AUCs of E2, FSH, LH, and F2iso,

$$\hat{\Omega} = (0.7132, 0.5347, 0.5318, 0.5218).$$

In this example, we chose F2iso as a reference biomarker. Note that the results remain the same when we use either biomarker as the reference. We calculated differences between individual markers and the reference marker of $\hat{\Omega}_c$ of (0.1915, 0.0130, 0.0100). The variance–covariance matrix $\hat{\Sigma}_c$ of $\hat{\Omega}_c$ was further estimated by (3):

$$\hat{\Sigma}_c = \begin{pmatrix} 0.0080 & 0.0033 & 0.0028 \\ 0.0033 & 0.0136 & 0.0095 \\ 0.0028 & 0.0095 & 0.0078 \end{pmatrix}.$$

The chi-square statistic was calculated, $T_c = \hat{\Omega}_c (\hat{\Sigma}_c)^{-1} \hat{\Omega}_c' = 5.06$, under H_0 . The p -value given by $P(\chi_3^2 \geq 10.85) = 0.17$ for this test does not show a significant difference among these markers based on their AUCs.

As these markers were developed to screen a large population of women and distinguish between ovulatory and anovulatory menstrual cycles, the accuracy at a FPR less than 0.6 was also determined to be important. In fact, similar decisions to look at small FPRs are often made for cancer screening markers. We conducted a homogeneity test of all these markers on the basis of the pAUCs with $\alpha = 0.05$. We calculated the partial AUC estimates of the biomarkers, (0.3590, 0.2286, 0.2196, 0.2026). This gave differences between individual markers and the reference marker of (0.1564, 0.0260, 0.0170). The variance–covariance matrix $\hat{\Sigma}_c$ of $\hat{\Omega}_c$ was further estimated by (3):

$$\begin{pmatrix} 0.0047 & 0.0026 & 0.0025 \\ 0.0026 & 0.0043 & 0.0030 \\ 0.0025 & 0.0030 & 0.0029 \end{pmatrix}.$$

The chi-square statistic was $T_c = \hat{\Omega}_c (\hat{\Sigma}_c)^{-1} \hat{\Omega}_c' = 8.14$ under H_0 . The p -value is 0.04 for this test, indicating a significant difference among these markers based on their partial AUCs.

The results we derived in Theorem 1 were used for power analysis. Suppose the variance–covariance matrix of $\hat{\Omega}$ remains the same under H_a . With a hypothesized Ω_c and type I error rate 0.05, the power under H_a is given by

$$P(X^2 > \chi_{3,0.05}^2),$$

where $X^2 \sim \chi_{3,\varphi}^2(\hat{\phi})$, and $\chi_{3,\alpha}^2$ is the upper α critical value of a chi-square distribution with three degrees of freedom. Here, φ is the estimate of the non-centrality parameter,

$\hat{\phi} = \Omega_c' \sum_c \hat{\Omega}_c$. Table I gives the powers for various hypothesized differences between partial AUCs of the reference marker and the other markers.

We also illustrate the sample size calculation using the method proposed in Section 2.2. Let the hypothesized difference between partial AUCs of the reference marker and the other markers be (0.15, 0.05, 0.05). The power is 44.49% in Table I with 35 anovulatory women and 224 ovulatory women in the original study. We calculated the required sample sizes to obtain 80% power. With type I error rate 0.05, the non-centrality parameter was calculated to be $\varphi_{0.05,0.80} \approx 14.64$. We then used the function *uniroot* in R package to solve the equation (6) and obtained the required sample sizes, $I = 96$ and $J = 48$, with the ratio of 2:1 between anovulatory and ovulatory women. Comparing with the original sample sizes, we can see that the required sample size for ovulatory women is only one-fourth of the original sample size. But the power of 80% is almost twice as large as the original power. Interestingly, the findings indicate that recruiting less anovulatory women and more ovulatory women may dramatically increase the test power.

Because a significant difference among pAUCs was found, we conducted post hoc pairwise comparisons while using the conservative Bonferroni criteria to adjust for multiple comparisons. The estimated pairwise differences between the reference marker and other markers are given by $\hat{\theta}_\ell$ for $\ell = 1, \dots, 3$, and the estimated variances of these differences are given by the diagonal elements in $\hat{\Sigma}_c$. We may also estimate the variances using results in Theorem 2. In fact, variance estimates given by Theorem 2 are similar to those in $\hat{\Sigma}_c$. As we are interested in whether other markers have different accuracy from the reference marker, the z -statistics for pairwise comparisons are (2.2785, 0.3972, 0.3140). The Bonferroni adjusted threshold is given by $z_{0.05/3} = 2.12$. Comparing the z -statistics with this threshold, it can be seen that E2 has different accuracy from F2isoprostane in distinguishing women with and without ovulation, and there is no significant difference between other markers and F2isoprostane.

5. A simulation study

We report simulation studies to evaluate the simulated type I error rates of the proposed test procedure. We simulated 1000 datasets under multivariate normal and multivariate log-normal distributions. We simulated multivariate normal random variables $\mathbf{X} \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\mathbf{Y} \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$, where $\boldsymbol{\mu}_X = (1, 1, 2, 2, 1.5, 1.5)$, $\boldsymbol{\mu}_Y = (0, \dots, 0)$, and $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_Y$ is the variance–covariance matrix with diagonal elements (1, 1, 4, 4, 2.25, 2.25) and correlation

coefficient, ρ . Here, ρ gives within-subject correlation. We let $L = 3$ markers and $K = 2$ repeated measurements for each subject. We chose the correlation coefficient ρ to be $-0.2, -0.1, 0, 0.2$, or 0.5 . We let $I = J = (50, 100, 200)$. For comparing AUCs, we let $W(u) = 1$ for $0 < u < 1$. For comparing partial AUCs, we let $W(u) = 1$ for $0 < u < 0.8$. To simulate multivariate log-normal random variables, we applied exponential transformation, $\mathbf{X} = \exp(\tilde{\mathbf{X}})$ and $\mathbf{Y} = \exp(\tilde{\mathbf{Y}})$, to get simulated log-normal data, where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ were simulated using the aforementioned settings.

We also simulated multivariate log-normal random variables under unequal sample sizes and unequal covariance matrices with $L = 3$ markers and $K = 2$ repeated measurements for each subject. In the simulation, we let $I = 80$, and $J = (100, 200)$. We let $\boldsymbol{\mu}_X = (0, \dots, 0)$, $\boldsymbol{\mu}_Y = (0, \dots, 0)$, and the covariance matrix for the non-diseased population have diagonal elements $(1, 1, 4, 4, 2.25, 2.25)$ and a correlation coefficient, ρ , and let the covariance matrix for the non-diseased population have the same diagonal elements but with a different correlation coefficient $\rho + 0.2$. We again let ρ be $-0.2, -0.1, 0, 0.2$, or 0.5 . We first simulated multivariate normal random variables $\tilde{\mathbf{X}} \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\tilde{\mathbf{Y}} \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$. The exponential transformation, $\mathbf{X} = \exp(\tilde{\mathbf{X}})$ and $\mathbf{Y} = \exp(\tilde{\mathbf{Y}})$, was then applied to get simulated log-normal data.

It is clear that the null hypothesis of equal AUCs or equal partial AUCs is true under these simulation settings. We then applied the proposed simultaneous comparison procedure to simulated datasets. For each simulated dataset, we estimated $\hat{\boldsymbol{\Omega}}_c = (\hat{\Omega}_2 - \hat{\Omega}_1, \hat{\Omega}_3 - \hat{\Omega}_1)$ and its variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_c$. We then calculated the χ^2 statistic, $T_c := \hat{\boldsymbol{\Omega}}_c \sum_c^{-1} \hat{\boldsymbol{\Omega}}_c'$. The null hypothesis was rejected if $T_c > \chi^2_2(0.95)$. We counted the frequency that the null hypothesis of equal AUCs or pAUCs was rejected out of 1000 simulated datasets in each simulation setting. Tables II and III show simulated rejection rates of our test procedure with the nominal level of 0.05. For sample sizes of 100 and 200, most rejection rates are within the 95% prediction interval.

6. Discussion

This article provides formal methods for the homogeneity test of several markers. The proposed methods are nonparametric and have important applications in comparing diagnostic markers. The associated theoretical properties are easy to implement in clustered ROC data. They provide new insights for analyzing clustered ROC data. Currently, the existing statistical inferences for evaluating complex ROC markers mainly rely on bootstrap or other resampling methods. The results studied in this paper give the closed-form expression of covariance structure for both within and between empirical ROC curves for clustered data. We provided the formula for power calculation and the numerical method for sample size calculation. The R code for conducting the proposed tests is available from the first author.

Acknowledgments

The authors thank anonymous referees, the associate editor, and the editor for their constructive comments and useful suggestions. The work was supported in part with funding from the American Chemistry Council and the

Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The project described here was supported by Award Number R15CA150698 from the National Cancer Institute under the American Recovery and Reinvestment Act of 2009, by Award Number H98230-11-1-0196 from the National Security Agency, and by Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service, project grant #XVA 61-036.

References

1. Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics*. 2000; 56 (4):1082–1087. [PubMed: 11129464]
2. McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics*. 2002; 58(3):657–664. [PubMed: 12230001]
3. Wactawski-Wende J, Schisterman EF, Hovey KM, Howards PP, Browne RW, Hediger M, Liu A, Trevisan M. BioCycle Study Grp. BioCycle Study: design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. *Paediatric and Perinatal Epidemiology*. 2009; 23(2):171–184. [PubMed: 19159403]
4. Schisterman EF, Gaskins AJ, Mumford SL, Browne RW, Yeung E, Trevisan M, Hediger M, Zhang C, Perkins NJ, Hovey K, et al. Influence of endogenous reproductive hormones on F-2-isoprostane levels in premenopausal women. *American Journal of Epidemiology*. 2010; 172(4):430–439. [PubMed: 20679069]
5. Zou K, Hall W, Shapiro D. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*. 1997; 16(19):2143–2156. [PubMed: 9330425]
6. DeLong ER, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
7. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997; 53:567–578. [PubMed: 9192452]
8. Zou KH. Comparison of correlated receiver operating characteristic curves derived from repeated diagnostic test data. *Academic Radiology*. 2001; 8(3):225–233. [PubMed: 11249086]
9. Hochberg, Y.; Tamhane, AC. *Multiple Comparison Procedures*. Wiley; New York: 1987.
10. Cai T, Pepe M. Semi-parametric ROC analysis to evaluate biomarkers for disease. *Journal of The American Statistical Association*. 2002; 97:1099–1107.
11. Silverman, BW. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC; London: 1986.
12. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989; 76:585–592.
13. 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]
14. Serfling, RJ. *Approximation Theorems of Mathematical Statistics*. Wiley; New York: 1980.
15. Hsieh F, Turnbull BW. Non- & semi- parametric estimation of the receiver operating characteristics (ROC) curve. *Annals of Statistics*. 1996; 24:25–40.
16. Van der Vaart, A.; Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag; New York: 1996.

Appendix: Proofs

We follow similar lines in the proof of Theorem 1 in [15] and use continuous mapping theorem [16, Theorem 1.3.6,] to show that $\sqrt{I}(\hat{\Omega}_c - \Omega_c)$ is asymptotically equivalent to

$$\sum_{i=1}^I \tilde{\mathbf{W}}_i / \sqrt{I} - \lambda \sum_{j=1}^J \tilde{\mathbf{V}}_j' / \sqrt{J}. \quad (7)$$

Note that the two summands of (7) are independent. Also, the first summation of (7) is the sum of independent mean zero random vectors, $\tilde{\mathbf{W}}_i$, and the second summation of (7) is the sum of independent mean zero random vectors, $\tilde{\mathbf{V}}_j$. Thus, it follows from multivariate central limit theorem [14, Theorem 1.9.2B] that

$$\sqrt{I}(\hat{\boldsymbol{\Omega}}_c - \boldsymbol{\Omega}_c) \xrightarrow{d} N_L(0, \sum_c = E(\tilde{\mathbf{W}}_i \tilde{\mathbf{W}}_i') + \lambda E(\tilde{\mathbf{V}}_j \tilde{\mathbf{V}}_j')). \quad (8)$$

The eigen-decomposition of $\boldsymbol{\Sigma}_c$ is given by $\sum_{\ell=1}^{L-1} \lambda_{(\ell)} \mathbf{Q}_\ell \mathbf{Q}_\ell'$, where $\lambda_{(\ell)}$ is the ℓ th largest eigenvalue out of $L - 1$ eigenvalues, and $\mathbf{Q}_\ell$ is the corresponding orthonormal eigenvector. On the basis of (8), we have

$$\left\| I \sum_c - I \sum_c \right\| \xrightarrow{p} 0. \quad (9)$$

where $\|\cdot\|$ is a norm on $R^{(L-1)^2}$. Denote

$$Z_\ell = \lambda_{(\ell)}^{-1/2} \hat{\boldsymbol{\Omega}}_{c,\ell} \mathbf{Q}_\ell.$$

$\ell = 1, \dots, L - 1$. Using Cramer–Wold device, we can show that asymptotically, Z_ℓ 's are independent and follow standard normal. We substitute $\hat{\lambda}_{(\ell)}$ for $\lambda_{(\ell)}$ and $\hat{\mathbf{Q}}_\ell$ for $\mathbf{Q}_\ell$ in Z_ℓ . Using the law of large numbers, it follows from (9) that

$$(\hat{Z}_1, \hat{Z}_2, \dots, \hat{Z}_{L-1}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_{L-1}).$$

Because T_c is essentially $\sum_{\ell=1}^{L-1} \hat{Z}_\ell^2$, the proof of Theorem 1 is completed by applying Theorem 3.5 in [14].

Figure 1.
Relative frequency histograms of ovulatory and anovulatory women for the markers estradiol (E2), follicle stimulating hormone (FSH), luteinizing hormone (LH), and F2-isoprostanes (F2iso).

Figure 2.

Empirical ROC curves and their derivative functions of the markers estradiol (E2), follicle stimulating hormone (FSH), luteinizing hormone (LH), and F2-isoprostanes (F2iso).

Table I

Power of the homogeneity test.

(μ_1, μ_2, μ_3)	Power (%)
(0.15, 0.05, 0.05)	44.49
(0.15, 0.10, 0.05)	60.27
(0.15, 0.10, 0.10)	62.49
(0.20, 0.05, 0.05)	41.87
(0.20, 0.10, 0.05)	85.25

Table II

Rejection rates of the proposed procedure for equal sample sizes.

ρ	Multivariate normal					
	AUC			pAUC		
	$I = J = 50$ (%)	100 (%)	200 (%)	$I = J = 50$ (%)	100 (%)	200 (%)
-0.2	7.90	6.20	5.80	8.20	6.50	6.00
-0.1	6.80	6.20	6.20	6.80	6.80	6.20
0	6.40	5.90	6.20	6.70	5.70	5.90
0.2	8.20	5.80	5.80	8.30	5.70	5.80
0.5	5.60	5.30	5.20	5.90	5.50	5.20

Multivariate log-normal					
-0.2	7.10	5.00	5.70	6.90	4.90
-0.1	6.90	4.70	4.90	7.20	5.10
0	6.90	6.10	6.00	7.20	5.80
0.2	7.50	6.10	5.80	7.20	5.60
0.5	6.00	6.80	5.80	5.90	4.70

The 95% predictive interval for the error rate is (3.5%, 6.5%).

Table III

Rejection rates for unequal sample sizes with $I = 80$.

ρ	AUC		pAUC	
	$J = 100$ (%)	$J = 200$ (%)	$J = 100$ (%)	$J = 200$ (%)
-0.2	5.40	5.60	7.10	5.80
-0.1	5.60	6.30	6.10	5.60
0	5.50	5.40	5.30	4.00
0.2	7.60	5.50	5.70	5.20
0.5	3.70	5.70	3.80	3.80

The 95% predictive interval for the error rate is (3.5%, 6.5%).