Published in final edited form as:

Stat Methods Med Res. 2018 October ; 27(10): 2933–2945. doi:10.1177/0962280216689806.

Biomarker validation with an imperfect reference: Issues and bounds

Sarah C Emerson¹, Sushrut S Waikar², Claudio Fuentes¹, Joseph V Bonventre², and Rebecca A Betensky³

¹Department of Statistics, Oregon State University, Corvallis, OR, USA

²Renal Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

³Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Abstract

Motivated by the goal of evaluating a biomarker for acute kidney injury, we consider the problem of assessing operating characteristics for a new biomarker when a true gold standard for disease status is unavailable. In this case, the biomarker is typically compared to another imperfect reference test, and this comparison is used to estimate the performance of the new biomarker. However, errors made by the reference test can bias assessment of the new test. Analysis methods like latent class analysis have been proposed to address this issue, generally employing some strong and unverifiable assumptions regarding the relationship between the new biomarker and the reference test. We investigate the conditional independence assumption that is present in many such approaches and show that for a given set of observed data, conditional independence is only possible for a restricted range of disease prevalence values. We explore the information content of the comparison between the new biomarker and the reference test, and give bounds for the true sensitivity and specificity of the new test when operating characteristics for the reference test are known. We demonstrate that in some cases these bounds may be tight enough to provide useful information, but in other cases these bounds may be quite wide.

Keywords

Biomarkers; conditional independence; diagnostic tests; imperfect reference; sensitivity; specificity

1 Introduction

Diagnosis of many medical conditions and diseases depends upon the use of diagnostic tests or biomarkers. In the development of such tests or biomarkers, it is necessary to assess the

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Sarah C Emerson, Department of Statistics, Oregon State University, 239 Weniger Hall, Corvallis, OR 97331, USA. emersosa@stat.oregonstate.edu.

Declaration of conflicting interests

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

performance and reliability of these biomarkers. Our initial motivating problem was to quantify the accuracy of a recently developed biomarker for acute kidney injury (AKI). The currently used tests for AKI depend upon the serum creatinine level, which is widely acknowledged to be an inferior diagnostic tool for many cases of AKI.¹ Unfortunately, the true presence or absence of AKI is difficult to establish without invasive and potentially risky histopathological assessment. Thus, in most study settings we are left without knowledge of the true classification to which we may compare a new biomarker. In many such cases where the true reference standard is unavailable, researches employ comparisons between the new test and an imperfect reference test, or imperfect gold standard. As discussed in Waikar et al.,² this can lead to dramatically incorrect conclusions about the performance of the new test if the imperfect reference test is treated as a true gold standard. Therefore, it is important to use information about the performance and fallibility of the new test in using such comparisons.

A number of authors have proposed approaches to using the information contained in the comparisons between a new diagnostic test and an imperfect reference test. These approaches generally require that the operating characteristics of the reference test be known or that additional reference tests be available, and furthermore, most methods incorporate strong assumptions about the dependence relationship between the two tests. One such common assumption is conditional independence, i.e., that the reference test and new test are independent given disease status. This is a very strong and often implausible assumption that we will explore in greater depth in the following section.

Employing the conditional independence assumption, Staquet et al.³ derive expressions for the operating characteristics of a new test under several different settings, and Walter and Irwig⁴ and Pepe and Janes⁵ develop and discuss latent class analysis (LCA). Yang and Becker⁶ addressed lack of conditional independence with multivariate marginal models when multiple tests are compared. While removing the constraint of the conditional independence assumption, the marginal models nevertheless must incorporate some restrictive dependence models in order to be identifiable. The choice of which interactions to include in the model may significantly change the resulting estimates. Other approaches to correcting the errors that result from an imperfect reference test include discrepant resolution (DR), a method that involves using a third test to resolve the cases for which the new test and reference test disagree and is now considered disfavored due to bias and dependence of the combined reference standard on the new test^{7,8}; and the composite reference standard (CRS) of Alonzo and Pepe⁷ which creates a reference standard from a combination of two reference tests. These two approaches seek to reduce the bias in the estimates of the operating characteristics of the new test by improving the performance of the reference to which the new test is compared, and therefore the success of both methods depends greatly upon the reliability of the available reference tests.

The goal of this work is to examine the extent to which comparisons of a new test to an imperfect reference test can provide information about the operating characteristics of the new test. Similar investigations have been performed in Vacek,⁹ Torrance-Rynard and Walter,¹⁰ Spencer¹¹ and Walter et al.¹² The primary contribution of this current discussion is to give precise bounds on the *true* operating characteristics of the new test, as a function of

the operating characteristics of the reference test and the data comparing the two tests. We make explicit the exact form of the bounds on these quantities of interest.

2 Background and problem statement

Here, we consider some of the issues that arise in validating a new test when it is difficult to establish the true disease status, and so an imperfect reference biomarker is used as a surrogate to which a new test may be compared. We begin with a review of some of the vocabulary and notation. The true presence or absence of the disease would be established via a *reference standard*: the current best definition of the disease. We will denote the disease by D , so that $D_i=1$ if patient i has the disease and $D_i=0$ if patient i is healthy. Unfortunately, in many cases the reference standard may be difficult or impossible to obtain. The reference standard is not to be confused with the *reference test* or, equivalently, imperfect gold standard, which we will denote by R . The reference test R is a test that is currently used for diagnosing the presence/absence of disease D , but it may be subject to diagnostic error. If patient i has a positive test result for test R (where positive is taken to mean indicating the presence of the disease condition), then $R_i=1$; otherwise $R_i=0$. Similarly, we use T to denote the new test under consideration, and let $T_i=1$ if patient i has a positive result according to test T , otherwise $T_i=0$. Frequently, the tests R and T are based on quantitative measures that are converted into qualitative (generally +/- or, equivalently, 1/0) results.

A general study framework to gauge the accuracy of the new test T consists of performing both tests T and R on the same sample of subjects from the target population. The results of the two tests are compared in a 2×2 table that summarizes the degree of agreement between the two tests:

	$R = 1$	$R = 0$
$T = 1$	n_{11}	n_{10}
$T = 0$	n_{01}	n_{00}

Here, n_{ij} is the number of subjects for which $T=i$ and $R=j$. These counts, divided by the total number of subjects in the sample, provide estimates of the probabilities

$$p_{11} = P(T = 1, R = 1) \quad p_{10} = P(T = 1, R = 0) \\ p_{01} = P(T = 0, R = 1) \quad p_{00} = P(T = 0, R = 0)$$

From the observed table, we can obtain estimates such as the *apparent sensitivity* a_T and *apparent specificity* b_T of the new test T based on comparisons to R

$$a_T = P(T = 1 \mid R = 1) \approx \frac{n_{11}}{n_{11} + n_{01}}$$

$$b_T = P(T = 0 \mid R = 0) \approx \frac{n_{00}}{n_{10} + n_{00}}$$

Frequently, some information about the performance of R is used to interpret comparisons between R and T , or at the very least it is assumed that R is reasonably accurate in diagnosing the disease. In some cases, more than one reference test may be used for comparison, which may provide more information if certain assumptions hold, as we will discuss later.

The goal of making comparisons between R and T is to assess the diagnostic accuracy of the new test T . Diagnostic accuracy is defined as “the extent of agreement between the outcome of the new test and the reference standard”⁸ when a reference standard exists; generally, diagnostic accuracy refers to the ability of a test to distinguish between the presence and absence of the disease. The most common measures of diagnostic accuracy are the sensitivity and specificity of the test

$$\text{Sensitivity of test } T = \alpha_T = P(T = 1 \mid D = 1)$$

$$\text{Specificity of test } T = \beta_T = P(T = 0 \mid D = 0)$$

The positive predictive value (PPV) and negative predictive value (NPV) of the test are also important measures of test performance, and in fact are frequently the more useful indicators of the diagnostic merit of the test. They are defined as

$$\text{PPV of test } T = P(D = 1 \mid T = 1)$$

$$\text{NPV of test } T = P(D = 0 \mid T = 0)$$

PPV and NPV are functions of the sensitivity and specificity of the test, as well as the disease prevalence $\theta = P(D = 1)$ in the target population.

When true disease status is not known, comparisons between R and T are used to estimate the operating characteristics of T . However, these comparisons are not fully informative about the operating characteristics of T . Full information for this problem would consist of a $2 \times 2 \times 2$ table comparing the results of R and T and the true disease status D

$D = 1$			$D = 0$		
	$R = 1$	$R = 0$		$R = 1$	$R = 0$
$T = 1$	n_{111}	n_{101}	$T = 1$	n_{110}	n_{100}
$T = 0$	n_{011}	n_{001}	$T = 0$	n_{010}	n_{000}

The quantities of interest can be obtained from the two-dimensional marginal 2×2 table comparing T to D , but we have only observed the two-dimensional marginal 2×2 table comparing T to R . If information about the operating characteristics of R relative to D is available, then we also have the marginal 2×2 table comparing R to D . Unfortunately, even the knowledge of two of the 2×2 marginal tables is not sufficient to fill in the full table, as has been extensively studied in the context of the ecological fallacy problem.^{13–15} The most we can do with these marginal tables is to place bounds on the contents of the full table, which we discuss in following sections. The bounds imposed by the margins are well known for a two-dimensional contingency table in the ecological fallacy setting, as developed in Duncan and Davis.¹⁶ We develop similar bounds for the three-dimensional context discussed here, and we extend to bounds on the operating characteristics of the new test T . We emphasize the important but often neglected point that without making further untestable assumptions about the relationships between the tests, this is the most that can be learned from the observed data.

3 Implausibility of conditional independence

3.1 Definition of conditional independence

Conditional independence between the tests means that given disease status, the results of the reference test and the new test are independent. In the context employed here, this means that the qualitative test results ($\pm$ or $1/0$) are independent. Note that this is not necessarily equivalent to the assumption that the quantitative measures that may underlie these tests are conditionally independent given disease status. Conditional independence means that, for $d=0, 1$

$$P(R = 1, T = 1 \mid D = d) = P(R = 1 \mid D = d)P(T = 1 \mid D = d)$$

If the two tests are physiologically related, it is widely acknowledged that conditional independence is unlikely to hold.^{5,6,17} Likewise, it has been well established that even if the two tests are measuring physiologically distinct mechanisms or quantities, the conditional independence assumption will generally be violated if there is any dependence of the test results on disease severity. Here, we further demonstrate that the assumption of conditional independence constrains the possible values of disease prevalence, $\theta = P(D = 1)$.

3.2 Disease prevalence and conditional independence

A further drawback to the conditional independence assumption is the fact that for a given contingency table comparing R and T , conditional independence may not be possible at a given value of the disease prevalence, $\theta = P(D = 1)$. To see this, suppose that the true cell probabilities in the contingency table comparing R and T are

S	$R = 1$	$R = 0$
$T = 1$	$a = P(T = 1, R = 1)$	$b = P(T = 1, R = 0)$

S	R = 1	R = 0
T = 0	$c = P(T=0, R=1)$	$d = P(T=0, R=0)$

Then, conditional independence implies that this table can be split into two tables

	D = 1		D = 0	
	R = 1	R = 0	R = 1	R = 0
T = 1	aa	βb	$(1 - \alpha)a$	$(1 - \beta)b$
T = 0	γc	δd	$(1 - \gamma)c$	$(1 - \delta)d$

for $\alpha, \beta, \gamma, \delta \in [0, 1]$ that satisfy the following constraints

$$\alpha a \delta d = \beta b \gamma c \text{ and } (1 - \alpha)a(1 - \delta)d = (1 - \beta)b(1 - \gamma)c$$

Then, $\theta = P(D=1) = aa + \beta b + \gamma c + \delta d$, and so the maximum possible value of θ is obtained when α, β, γ , and δ are all maximized subject to the constraints and similarly, the minimum possible value of θ occurs when α, β, γ , and δ are minimized subject to the constraints. Manipulating these constraints (see Appendix 1) and letting $r = bc/ad$, we obtain two potentially overlapping intervals for possible values of θ

$$\theta_U \in [(1 - r)a, a + b + c + rd]$$

$$\theta_L \in [(1 - r)d, ra + b + c + d]$$

if $ad > bc$, and

$$\theta_U \in [(1 - r)b, a + b + rc + d]$$

$$\theta_L \in [(1 - r)c, a + rb + c + d]$$

if $ad < bc$. These intervals may overlap or not, depending upon the values of a, b, c , and d . So for example, if the following table is observed

	R = 1	R = 0
T = 1	$a = 0.30$	$b = 0.05$
T = 0	$c = 0.15$	$d = 0.50$

and we ignore sampling variability, then for any $\theta \notin [0.285, 0.715]$, conditional independence is not possible. This phenomenon can produce nonsensical operating characteristic estimates for test T if the formulae of Staquet et al.³ are used. For instance, in

the above example, if test R is known to have sensitivity $\alpha_R = 0.9$ and specificity $\beta_R = 0.7$, we can use the observed proportion of positive results for test R to calculate the disease prevalence as $\hat{\theta} = 0.25$, which is outside of the range for which conditional independence is possible. Using these values of α_R and β_R , formula (8) on page 603 of Staquet et al.³ for the sensitivity of the new test produces 1.3, which is clearly an invalid value.

These considerations indicate that, in general, conditional independence given disease status may be an inappropriately strong and invalid assumption, even when the two tests are based on different and seemingly unrelated phenomena. As we will discuss in the following sections, and as has been noted by many other authors, assuming conditional independence when it does not in fact hold can lead to markedly inaccurate estimates of the operating characteristics for the new test.

4 Bounds on true sensitivity and specificity of a new biomarker

Given operating characteristics for the reference test R , along with the observed 2×2 table comparing the results of R and the new test T for the same set of subjects, we can derive bounds for the true sensitivity and specificity of T . This corresponds to using the marginal 2×2 table comparing R to D and the marginal 2×2 table comparing R to T for the purpose of learning about the 2×2 table comparing T to D . The bounds that we first present here do not take into account sampling variability; only variability due to differences in dependence structure is represented in our results. The sampling variability of the probabilities estimated from the observed table will of course depend on the number of subjects in the sample and has been extensively studied.¹⁸ Variability of the bounds we derive as a function of sampling variability is very difficult to express in closed form; we therefore suggest a bootstrap approach to estimating the variability of the bounds presented in this section, detailed in the subsequent subsection.

4.1 Operating characteristic bounds

Extending the definitions in Section 1, we define the following parameters:

- $\psi_R = P(D = 1 | R = 1)$, PPV of R ;
- $\eta_R = P(D = 0 | R = 0)$, NPV of R ;
- $a_T = P(T = 1 | R = 1)$, apparent sensitivity of T ;
- $b_T = P(T = 0 | R = 0)$, apparent specificity of T ;
- $\alpha_T = P(T = 1 | D = 1)$, true sensitivity of T ;
- $\beta_T = P(T = 0 | D = 0)$, true specificity of T ;
- $\pi_R = P(R = 1)$, positive probability of R ;
- $\tau_+ = P(D = 1, T = 1 | R = 1)$, positive concordance parameter;
- $\tau_- = P(D = 0, T = 0 | R = 0)$, negative concordance parameter.

The parameters ψ_R and η_R may be determined from the R vs. D table, which may be constructed from the operating characteristics of R ($\alpha_R = P(R = 1 | D = 1)$ and $\beta_R = P(R = 0 | D = 0)$) and the positive probability of R , π_R

$$\begin{aligned}\pi_R &= P(R = 1) = \alpha_R \theta + (1 - \beta_R)(1 - \theta) \Rightarrow \theta = P(D = 1) = \frac{\pi_R + \beta_R - 1}{\alpha_R + \beta_R - 1} \\ \psi_R &= P(D = 1 | R = 1) = \frac{P(R = 1 | D = 1)P(D = 1)}{P(R = 1)} = \frac{\alpha_R \theta}{\pi_R} \\ \eta_R &= P(D = 0 | R = 0) = \frac{P(R = 0 | D = 0)P(D = 0)}{P(R = 0)} = \frac{\beta_R(1 - \theta)}{(1 - \pi_R)}\end{aligned}$$

Then, using this notation, the following table is obtained:

$R = 1$	
$D = 1, T = 1$	$\tau_+ \pi_R$
$D = 1, T = 0$	$(\psi_R - \tau_+) \pi_R$
$D = 0, T = 1$	$(a_T - \tau_+) \pi_R$
$D = 0, T = 0$	$(1 - \psi_R - a_T + \tau_+) \pi_R$
$R = 0$	
$D = 1, T = 1$	$(1 - \eta_R - b_T + \tau_-)(1 - \pi_R)$
$D = 1, T = 0$	$(b_T - \tau_-)(1 - \pi_R)$
$D = 0, T = 1$	$(\eta_R - \tau_-)(1 - \pi_R)$
$D = 0, T = 0$	$(\tau_-)(1 - \pi_R)$

Examining the above table and using the fact that the probability in each cell must be between 0 and 1, we see that the positive and negative concordance parameters (τ_+ and τ_-) must satisfy the following bounds

$$\begin{aligned}\max(0, \psi_R + a_T - 1) &\leq \tau_+ \leq \min(\psi_R, a_T) \\ \max(0, \eta_R + b_T - 1) &\leq \tau_- \leq \min(\eta_R, b_T)\end{aligned}$$

Then, in terms of the known parameters for the test R , the observed R vs. T comparisons, and the concordance parameters, expressions may be obtained for the true sensitivity and specificity of the new test, namely

$$\begin{aligned}\alpha_T &= \frac{\tau_+ \pi_R + (1 - \eta_R - b_T + \tau_-)(1 - \pi_R)}{\theta} \\ \beta_T &= \frac{(1 - \psi_R - a_T + \tau_+) \pi_R + \tau_- (1 - \pi_R)}{(1 - \theta)}\end{aligned}$$

It is easy to see that the largest possible values of α_T and β_T result when τ_+ and τ_- are maximized, and the smallest possible values of α_T and β_T result when τ_+ and τ_- are minimized. Thus, the bounds on τ_+ and τ_- produce maximum and minimum possible values for the sensitivity and specificity of the new biomarker.

As an example, suppose that the following concordance table is observed

	$R = 1$	$R = 0$
$T = 1$	25	5
$T = 0$	10	60

From this table, we find that $a_T = \frac{25}{35} = 0.714$ and $b_T = \frac{60}{65} = 0.923$. Suppose further that the operating characteristics of R are known to be $\alpha_R = 0.9$ and $\beta_R = 0.95$, so R is a reasonably reliable test. From these operating characteristics and from the observed proportion of positive results for R , we can compute the disease prevalence $\theta = \frac{.35 + .95 - 1}{.9 + .95 - 1} = 0.353$. Thus,

the PPV and NPV of R are $\psi_R = \frac{\alpha_R \theta}{\pi_R} = 0.908$ and $\eta_R = \frac{\beta_R (1 - \theta)}{(1 - \pi_R)} = 0.946$. Then, using the

bounds on τ_+ and τ_- from above, we find that the maximum possible values for the true sensitivity and specificity of T are

$$\alpha_{T(\max)} = \frac{0.714(0.35) + 0.054(0.65)}{0.353} = 0.807$$

$$\beta_{T(\max)} = \frac{0.092(0.35) + 0.923(0.65)}{0.647} = 0.977$$

and the minimum possible values are

$$\alpha_{T(\min)} = \frac{0.622(0.35) + 0(0.65)}{0.353} = 0.617$$

$$\beta_{T(\min)} = \frac{0(0.35) + 0.869(0.65)}{0.647} = 0.873$$

To summarize, for the given table, we find the following bounds on the sensitivity α_T and specificity β_T :

	Lower bound	Upper bound
Sensitivity α_T	0.617	0.807
Specificity β_T	0.873	0.977

These ranges are reasonably narrow and perhaps constitute useful information about the test T . However, suppose instead that the observed concordance table comparing R and T were

S	R = 1	R = 0
T = 1	10	20
T = 0	10	60

In this case, we would find the following bounds:

	Lower bound	Upper bound
Sensitivity α_T	0.333	0.667
Specificity β_T	0.707	0.779

However, these bounds are much less informative, and the sensitivity bound in particular ranges from a very poor test to a moderately fair test. For this example, the available data may not provide sufficient information to decide on the usefulness of the proposed new test T .

4.2 Variability of OC bound estimates

Since the OC bounds depend in a very complicated way upon the probabilities $\{p_{11}, p_{10}, p_{01}, p_{00}\}$, it is quite difficult to derive closed-form expression for the variability of the OC bounds. However, a reasonable estimate of variability may be obtained by bootstrapping/resampling. Briefly,

1. For each of B different resampling iterations ($b = 1, \dots, B$), do:
  - a. Resample n cases with replacement from the original n cases.
  - b. Form the observed table based on the resampled cases, and compute the estimated probabilities for this iteration, $\{\hat{p}_{11}^{(b)}, \hat{p}_{10}^{(b)}, \hat{p}_{01}^{(b)}, \hat{p}_{00}^{(b)}\}$ based on this observed table.
  - c. Compute the estimated bounds for this iteration: $\hat{\alpha}_T^{\min(b)}, \hat{\alpha}_T^{\max(b)}, \hat{\beta}_T^{\min(b)}$, and $\hat{\beta}_T^{\max(b)}$.
2. Estimate the sampling variability of each bound as

$$\begin{aligned}\widehat{SE}(\hat{\alpha}_T^{min}) &= \left[\frac{1}{B-1} \sum_{b=1}^B \left(\hat{\alpha}_T^{min}(b) - \hat{\alpha}_T^{min}(\cdot) \right)^2 \right]^{1/2} \\ \widehat{SE}(\hat{\alpha}_T^{max}) &= \left[\frac{1}{B-1} \sum_{b=1}^B \left(\hat{\alpha}_T^{max}(b) - \hat{\alpha}_T^{max}(\cdot) \right)^2 \right]^{1/2} \\ \widehat{SE}(\hat{\beta}_T^{min}) &= \left[\frac{1}{B-1} \sum_{b=1}^B \left(\hat{\beta}_T^{min}(b) - \hat{\beta}_T^{min}(\cdot) \right)^2 \right]^{1/2} \\ \widehat{SE}(\hat{\beta}_T^{max}) &= \left[\frac{1}{B-1} \sum_{b=1}^B \left(\hat{\beta}_T^{max}(b) - \hat{\beta}_T^{max}(\cdot) \right)^2 \right]^{1/2}\end{aligned}$$

where

$$\begin{aligned}\hat{\alpha}_T^{min}(\cdot) &= \frac{1}{B} \sum_{b=1}^B \hat{\alpha}_T^{min}(b) & \hat{\alpha}_T^{max}(\cdot) &= \frac{1}{B} \sum_{b=1}^B \hat{\alpha}_T^{max}(b) \\ \hat{\beta}_T^{min}(\cdot) &= \frac{1}{B} \sum_{b=1}^B \hat{\beta}_T^{min}(b) & \hat{\beta}_T^{max}(\cdot) &= \frac{1}{B} \sum_{b=1}^B \hat{\beta}_T^{max}(b)\end{aligned}$$

These resampling estimates are reasonably close in simulations.

5 Audiology tests example

The following example is taken from Pepe and Janes⁵ and consists of the results of three different audiology screening tests for a sample of $n=666$ subjects. For these data, the true target condition status for the subjects is known, and so we are able to compare the results of different analysis methods to the truth. We use this example to illustrate the wide range of operating characteristics that are possible for a given test, even when the operating characteristics for two comparison tests are known. We also demonstrate the potential drawbacks of methods that attempt to return single estimates of the operating characteristics. The table comparing the results of the three tests is given in Table 1. Since the true disease status is available for these subjects, we are able to recreate the full table which includes disease status. From this full table, we can calculate the following estimates of disease prevalence and the operating characteristics for the three tests: prevalence $\theta = 0.42$; sensitivities $\alpha_R = 0.66$, $\alpha_S = 0.62$, and $\alpha_T = 0.75$; and specificities $\beta_R = 0.60$, $\beta_S = 0.64$, and $\beta_T = 0.46$.

First, we present the results of LCA applied to these data, as reported in Pepe and Janes.⁵ LCA, such as the methods proposed by Walter and Irwig⁴ and Pepe and Janes,⁵ assumes conditional independence between the tests. Examining the full table, we see that this assumption is incorrect for this data, and thus, we can expect that the results of LCA will be inaccurate. The parameter estimates from the latent class table are: prevalence $\hat{\theta}(LCA) = 0.54$; sensitivities $\hat{\alpha}_R(LCA) = 0.84$, $\hat{\beta}_S(LCA) = 0.76$, and $\hat{\alpha}_T(LCA) = 0.90$; and specificities $\hat{\beta}_R(LCA) = 0.87$, $\hat{\beta}_S(LCA) = 0.87$, and $\hat{\beta}_T(LCA) = 0.69$. These estimates are all higher than the true values.

Next, we consider the use of DR, treating tests R and S as the reference and resolver tests, respectively, with the goal of estimating the operating characteristics of T . DR is discussed in Alonzo and Pepe,⁷ Hagdu,¹⁹ U.S. Food and Drug Administration,⁸ and produces assigned disease status for each subject based on a simple majority of the three tests. Then, the operating characteristics of test T may be computed using this assigned status. DR has been discredited^{7,8,19} due to the fact that the assigned disease status depends upon the results of the new test. Estimates of the operating characteristics of test T obtained using assigned disease status according to DR are: prevalence $\hat{\theta}(DR) = 0.54$; sensitivity $\hat{\alpha}_T(DR) = 0.92$; and specificity $\hat{\beta}_T(DR) = 0.72$.

The CRS, proposed by Alonzo and Pepe⁷ addresses the issues with the DR by assigning disease status based only on the reference tests R and S . A case is labeled disease positive if either of the reference tests produces a positive result; otherwise, the subject is assigned a disease negative status. Using the CRS approach may or may not give better performance than using a single reference test in terms of evaluating the performance of a new test; while the sensitivity of the CRS will be at least as high as the maximum sensitivity of the two tests, the specificity may be lower than the specificity of either test. The exact performance of this approach depends on the performance of each contributing reference test, as well as on the dependence relationship between the two tests.²⁰ The results of this approach applied to the audiology data are shown below. Using the CRS results as the truth, the following parameter estimates are obtained: prevalence $\hat{\theta}(CRS) = 0.63$; sensitivity $\hat{\alpha}_T(CRS) = 0.79$; and specificity $\hat{\beta}_T = 0.66$.

Finally, we explore the bounds on the possible values of sensitivity and specificity of T if the operating characteristics of the two reference tests are known. This case is slightly more complicated than the situation discussed in Section 4.1, as there are two reference tests with presumed known operating characteristics instead of just one. However, completely analogous bounds can be constructed in this more complicated case. To obtain the tables that result in the minimum and maximum operating characteristics, we fix the margins of the $R \times S \times T$ table, but then allocate observations to $D=0$ or $D=1$ in such a way as to minimize (respectively, maximize) the agreement between D and T , while maintaining the agreement between D , R , and S (that is, the total number of $D=1$ counts shared between two rows with the same values of R and S must be constant). This preserves the operating characteristics of R and S while allowing the operating characteristics of T to vary. Here, we demonstrate the full tables that would obtain the minimum and maximum bounds on the operating characteristics. The full table that produces the upper bound on the sensitivity and specificity of T is given in Table 2. This upper bound table produces the maximum possible values of sensitivity and specificity for T while maintaining the operating characteristics of R and S . These upper bounds are calculated to be: sensitivity $\alpha_T(max) = 1.00$; specificity $\beta_T(max) = 0.64$. Similarly, the full table that produces the lower bound on the sensitivity and specificity of T is given in Table 3. From this table, the lower bounds for the sensitivity and specificity of T are found to be: sensitivity $\alpha_T(min) = 0.45$; specificity $\beta_T(min) = 0.25$.

To summarize, Table 4 contains the estimates of the operating characteristics of T for each of the methods considered, along with bootstrap intervals based on 5000 resampled datasets for all methods except the LCA. The provided bootstrap intervals are the (0.025, 0.975)

quantiles of the bootstrap distribution. We note that in computing the bounds, we have assumed that information about the performance of the other two tests is available. If this information is not available, or not very accurate, the resulting bounds for T will be even wider. This comparison shows that even in the arguably ideal situation of knowing everything about the comparison tests R and S , very little can be learned about the performance of T . The LCA method that assumes conditional independence and estimates all parameters produces inaccurate estimates, and the DR and CRS methods, which implicitly assume near perfect reference tests, also provide quite unreliable results. We see from the width of the bootstrap intervals for all methods that the sample size in this study is not large enough to provide good precision on the resulting estimates, partly because some cell counts are quite small (for instance, $R = 0$, $S = 1$, $T = 0$ has only 29 observations, and $R = 1$, $S = 1$, $T = 0$ has only 27).

6 Discussion

The bounds developed in this paper may be viewed as informal measures of the information content that is present in a given set of comparisons between different diagnostic tests or biomarkers, when information about some of the tests being compared is available. Whereas Pepe and Janes⁵ argue that knowledge about the marginal parameters and accuracies of reference tests is not particularly useful, we find that in some cases, this marginal information can be valuably informative. As we have demonstrated here, the marginal information does provide some degree of constraint on the parameters of the new test, and so in many cases, there may be more information in the comparisons than Pepe and Janes⁵ suggest. In some instances, the resulting bounds might be quite tight, and therefore no further assumptions or methodology are necessary for the intended purpose of establishing the diagnostic accuracy of the new test in question. For example, in the extreme case of complete agreement between the new test and reference test, knowledge of the reference test operating characteristics is easily seen to be equivalent to knowledge about the new test. Relaxing this extreme case just a bit, it follows that if the tests agree to a great extent then the resulting bounds on the new test operating characteristics will be quite narrow. On the other hand, the FDA guidance⁸ contains the following statement:

In rare circumstances, it may be possible to estimate sensitivity and specificity without using a reference standard in the study. This may be reasonable, for example, when the sensitivity and specificity of the designated comparative method are well established from previous evaluations against a reference standard in similar subject populations. Further elaboration of this subject is beyond the scope of this document. Here too, FDA recommends you consult with a CDRH statistician before using this approach.

Yet, as we have seen in the examples we considered here, even when the reference test is quite highly sensitive and specific, there may be cases when comparisons to such a reference test still yield a surprisingly small quantity of information about the operating characteristics of the new test. Therefore, our primary recommendation is to examine the bounds that result from a given comparison to see if that comparison provides any useful information about the new test in question. In cases where the resulting bounds are narrow, perhaps such a

comparison will be sufficient to go forward with. However, in the many cases where the resulting bounds are wide, a different study design framework might be necessary in order to assess the accuracy or utility of the new test in question.

All of our discussion has presumed knowledge about the reference test that may in fact be lacking or imprecise. If exact operating characteristics for the reference test are not known, but crude estimates or ranges are available, these ranges may be used to construct bounds by computing bounds for each possible combination of values for the reference test, and then taking the minima and maxima of all of the resulting bounds on the operating characteristics of the new test. This approach may well result in bounds that are so wide as to be uninformative. Thus, if very little is known about the reference test performance, then it is clear that a comparison to such a reference test is a futile exercise and can provide no information about a new test.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the following grants: National Institutes of Health R01 CA075971, T32NS048005, and R33 DK074099.

References

1. Coca SG, Parikh CR. Urinary biomarkers for acute kidney injury: perspectives on translation. *Clin J Am Soc Nephrol.* 2008; 3:481–490. [PubMed: 18256377]
2. Waikar SS, Betensky RA, Emerson SC, et al. Imperfect gold standards for kidney injury biomarker evaluation. *J Am Soc Nephrol.* 2012; 23:13–21. [PubMed: 22021710]
3. Staquet M, Rozenzweig M, Lee YJ, et al. Methodology for the assessment of new dichotomous diagnostic tests. *J Chron Dis.* 1981; 34:599–610. [PubMed: 6458624]
4. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol.* 1988; 41:923–937. [PubMed: 3054000]
5. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostat.* 2007; 8:474–484.
6. Yang I, Becker MP. Latent variable modeling of diagnostic accuracy. *Biometrics.* 1997; 53:948–958. [PubMed: 9290225]
7. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med.* 1999; 18:2987–3003. [PubMed: 10544302]
8. US Food and Drug Administration. Guidance for industry and FDA staff: statistical guidance on reporting results from studies evaluating diagnostic tests (Document Number 1620). Silver Spring, MD: US Food and Drug Administration, Center for Devices and Radiological Health, Diagnostic Devices Branch; 2007.
9. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics.* 1985; 41:959–968. [PubMed: 3830260]
10. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. *Stat Med.* 1997; 16:2157–2175. [PubMed: 9330426]
11. Spencer BD. When do latent class models overstate accuracy for diagnostic and other classifiers in the absence of a gold standard? *Biometrics.* 2012; 68:559–566. [PubMed: 22017371]
12. Walter S, Macaskill P, Lord SJ, et al. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Stat Med.* 2012; 31:1129–1138. [PubMed: 22351623]

13. Robinson WS. Ecological correlations and the behavior of individuals. *Am Soc Rev.* 1950; 15:351–357.
14. Greenland S, Robins J. Invited commentary: ecologic studies-biases, misconceptions, and counterexamples. *Am J Epidemiol.* 1994; 139:747–760. [PubMed: 8178788]
15. Freedman DA. Technical Report. Statistics Department, University of California; Berkeley, USA: 1994. Ecological inference and the ecological fallacy.
16. Duncan OD, Davis B. An alternative to ecological correlation. *Am Soc Rev.* 1953; 18:665–666.
17. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Method Med Res.* 1998; 7:354–370.
18. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies I. Indices of agreement and their relation to prevalence. *Am J Epidemiol.* 1966; 83:586–592. [PubMed: 5932702]
19. Hagdu A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. *J Clin Epidemiol.* 1999; 52(12):1231–1237. [PubMed: 10580787]
20. Naaktgeboren CA, Bertens LC, van Smeden M, et al. Value of composite reference standards in diagnostic research. *Br Med J.* 2013; 347

Appendix 1

Derivation of prevalence bounds

We suppose that the true cell probabilities are given in the following table, and we seek to identify possible values of the disease prevalence parameter $\theta = P(D = 1)$ for which conditional independence could hold. Conditional independence for the table

	$R = 1$	$R = 0$
$T = 1$	$a = P(T = 1, R = 1)$	$b = P(T = 1, R = 0)$
$T = 0$	$c = P(T = 0, R = 1)$	$d = P(T = 0, R = 0)$

means that that the table can be split into two tables, one for disease positive and one for disease negative, as follows,

$D = 1$		
	$R = 1$	$R = 0$
$T = 1$	αa	βb
$T = 0$	γc	δd

$D = 0$		
	$R = 1$	$R = 0$
$T = 1$	$(1 - \alpha)a$	$(1 - \beta)b$
$T = 0$	$(1 - \gamma)c$	$(1 - \delta)d$

for $\alpha, \beta, \gamma, \in [0, 1]$ which satisfy

$$\alpha a \delta d = \beta b \gamma c \text{ and } (1 - \alpha)a(1 - \delta)d = (1 - \beta)b(1 - \gamma)c$$

Since $\theta = \alpha a + \beta b + \gamma c + \delta d$, θ is maximized when α , β , γ , and δ are maximized, and θ is minimized when α , β , γ , and δ are minimized, subject to the constraints. Assuming that none of a , b , c , d are zero, we can rearrange these constraints to get

$$\alpha \delta = \frac{bc}{ad} \beta \gamma \text{ and } (1 - \alpha)(1 - \delta) = \frac{bc}{ad} (1 - \beta)(1 - \gamma)$$

Letting $r = \frac{bc}{ad}$, these constraints become

$$\alpha \delta = r \beta \gamma \text{ and } (1 - \alpha)(1 - \delta) = r(1 - \beta)(1 - \gamma) \quad (1)$$

Without loss of generality, we will assume that $ad \geq bc$, so $0 \leq r \leq 1$ (if this is not the case, the following arguments will hold with α and δ interchanged for β and γ). Since $r = bc/ad \leq 1$, we may freely choose β and γ in $[0, 1]$ to obtain a viable solution. For a given choice of β and γ , we can then solve for α and δ using the two constraints (1)

$$\alpha = \frac{(1 - r) + r(\beta + \gamma) \pm \sqrt{(r - 1 - r(\beta + \gamma))^2 - 4r\beta\gamma}}{2}$$

$$\delta = 1 - \alpha - r(1 - \beta - \gamma)$$

We designate by α_U the value of α that results when we replace the “ $\pm$ ” in the expression for α with “+,” and δ_U the corresponding value of δ , and similarly α_L will be the value of α when we replace the “ $\pm$ ” with “−” and δ_L again the corresponding value of δ . Then, it can be shown that α and δ are both increasing functions of β and γ , and so are maximized when β and γ are both 1, and minimized when β and γ are both 0. This leads to two possible intervals for the prevalence parameter $\theta = \alpha a + \beta b + \gamma c + \delta d$: the interval when α_U and δ_U are used, and the interval when α_L and δ_L are used

$$\theta_U \in [(1 - r)a, a + b + c + rd]$$

$$\theta_L \in [(1 - r)d, ra + b + c + d]$$

If $r = \frac{bc}{ad} \geq 1$, following the same reasoning, we would instead get intervals

$$\theta_U \in [(1 - r)b, a + b + rc + d]$$

$$\theta_L \in [(1 - r)c, a + rb + c + d]$$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Observed concordance data with true disease status.

<i>R</i>	<i>S</i>	<i>T</i>	<i>D</i> = 0	<i>D</i> = 1	Count
0	0	0	124	38	162
0	0	1	58	27	85
0	1	0	23	6	29
0	1	1	28	22	50
1	0	0	20	11	31
1	0	1	47	28	75
1	1	0	13	14	27
1	1	1	76	131	207
Total			389	277	666

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2
Observed concordance data with OC upper bounds for T : fixed (known) OC values for R and S .

R	S	T	$D = 0$	$D = 1$	Count
0	0	0	162	0	162
0	0	1	20	65	85
0	1	0	29	0	29
0	1	1	22	28	50
1	0	0	31	0	31
1	0	1	36	39	75
1	1	0	27	0	27
1	1	1	62	145	207
Total					666

Table 3
Observed concordance data with OC lower bounds for T : fixed (known) OC values for R and S :

R	S	T	$D = 0$	$D = 1$	Count
0	0	0	97	65	162
0	0	1	85	0	85
0	1	0	1	28	29
0	1	1	50	0	50
1	0	0	0	31	31
1	0	1	67	8	75
1	1	0	0	27	27
1	1	1	89	118	207
Total			389	277	666

Table 4 Comparison of OC estimates for T .

	Prevalence θ	Sensitivity α_T	Specificity β_T
Truth	0.42	0.75	0.46
LCA	0.54	0.90	0.69
DR	0.54	0.92	0.72
(Bootstrap)	(0.50, 0.58)	(0.90, 0.95)	(0.67, 0.77)
CRS	0.63	0.79	0.66
(Bootstrap)	(0.59, 0.67)	(0.75, 0.83)	(0.59, 0.71)
Upper bound	0.42	1.00	0.64
(Bootstrap)	(0.29, 0.54)	(1.00, 1.00)	(0.58, 0.71)
Lower bound	(same	0.45	0.25
(Bootstrap)	as above)	(0.39, 0.53)	(0.20, 0.32)

LCA: latent class analysis; DR: discrepant resolution; CRS: composite reference standard.