ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2016)
65, *Part* 2, *pp.* 315–329

Comparing two binary diagnostic tests with repeated measurements

Stefan H. Steiner

University of Waterloo, Canada

Oana Danila

F. Hoffman–La Roche, Basel, Switzerland

and Cecilia A. Cotton, Daniel Severn and R. Jock Mackay

University of Waterloo, Canada

[Received October 2014. Final revision July 2015]

Summary. We compare two binary diagnostic tests when each subject is measured more than once with each test and with a gold standard. We introduce a new model that allows the correlation between two measurements on a single subject by the same test to be different from the correlation between two measurements by different tests. We show that moment estimators of the population parameters for the mean sensitivities and specificities are virtually identical to the maximum likelihood estimates from our random-effects model. We apply the model to data comparing two rapid malaria tests and provide guidance for choosing the number of subjects and repeated measurements.

Keywords: Diagnostic tests; Gold standard; Moment estimators; Random effects; Repeated testing; Sensitivity and specificity

1. Introduction

In medical applications, there are many diagnostic tests such as bacterial cultures, radiographic images and biochemical tests used to determine the disease status of a subject. Binary diagnostic tests are used to identify the presence or absence of a disease in subjects who have signs or symptoms. Breast cancer screening with mammography of women over 50 years of age and cervical cancer screening with a Pap smear are two examples. A key goal is to avoid false positive and false negative results. This paper addresses the comparison of the statistical properties of two such tests when a gold standard system is available to verify the disease status of each subject in the study and we have repeated measurements for each test on each subject.

The main contribution of the paper is to examine the advantages of making repeated measurements by each test when comparing the sensitivity and specificity. We also show how to plan and analyse the data from such an investigation. It may be less expensive to measure every subject multiple times with each test rather than to obtain more subjects, especially if the gold standard system is expensive or invasive. We quantify this trade-off when we consider planning a comparison study with the possibility of repeated measurements for each test.

Address for correspondence: Stefan H. Steiner, Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada.
E-mail: shsteiner@uwaterloo.ca

0035–9254/16/65315

Table 1. Malaria tests data from Hopkins *et al.* (2007)†

Number of positive HRP2 tests	*Malaria positive subjects for the following numbers of positive pLDH tests*			*Malaria negative subjects for the following numbers of positive pLDH tests*		
	0	*1*	*2*	*0*	*1*	*2*
0	11 (11.2)	1 (1.3)	3 (2.6)	579 (578.8)	1 (1.0)	0 (0.0)
1	4 (3.5)	1 (0.8)	2 (2.5)	9 (9.4)	0 (0.0)	0 (0.0)
2	18 (18.5)	10 (9.5)	239 (239.1)	40 (39.8)	0 (0.0)	0 (0.0)

†Interior values give numbers of subjects; values in parentheses give the expected number for the eGRE model discussed in Section 5.

Our motivation is the recommendation of Baker *et al.* (1991) to use replicate observations in observer agreement studies and the data collected by Hopkins *et al.* (2007) where the objective was to compare the sensitivity and specificity of two rapid malaria tests pLDH and HRP2. A total of 918 subjects were tested two times by each of the rapid tests. In addition, the malaria status of each subject was definitively determined by using microscopic examination of blood smears. The data are summarized in Table 1. A key feature of the motivating application is that each test was repeated twice on each subject. For example, we see in Table 1 that, of the 628 malaria negative subjects, 40 tested positively two times (out of two) with the HRP2 test but zero times (out of two) with the pLDH test. We explain the expected counts given in Table 1 in Section 4 where we revisit the motivating application.

We start with some basic assumptions. First, in common with Qu *et al.* (1996), Fujisawa and Izumi (2000) and Albert and Dodd (2004), we assume that the sensitivity (or specificity) for either test varies from subject to subject in the study population. Some subjects are more likely to be correctly diagnosed because there may be one or more latent covariates that affect the sensitivity or specificity. For instance, sensitivity could depend on the severity of the disease at the time of testing. Second, given the disease status, we expect that the results of two tests on the same subject will be dependent and, furthermore, the dependence between the results of measuring a subject twice with the same test will differ from the dependence between the results of measuring a subject once with each test. Third, we assume that, given the disease status for any subject and the subject test-specific sensitivity (or specificity), the results for repeated measurements on this subject by either test are statistically independent. We include these assumptions in the model that is described in Section 2. Albert and Dodd (2004, 2008), Qu *et al.* (1996) and Pepe (2003) made similar assumptions when considering the assessment of one or more diagnostic tests without repeated measurements by the same test.

Many researchers have considered the comparison of two diagnostic tests. See, for example, DeLong *et al.* (1988), Pepe (2003) and Zhou *et al.* (2011), all of whom used models based on receiver operating characteristic curves that discretize a continuous output that can be measured. Nofuentes and Del Castillo (2007) proposed a method for comparison of two tests based on a multinomial model. Biggerstaff (2000) used a graphical approach. None of these methods deals with multiple measurements by each test on each subject.

We have organized the paper as follows. In the next section, we describe a random-effects model that allows for varying sensitivities (and specificities) across subjects and induces dependences between the measurements made on each subject. In the following section, we define and determine the properties of some simple moment estimates of the population-average sensitivity (and specificity). We also discuss the estimation of the variance parameters of the random-effects distribution. Next, in Section 4, we use the methodology to analyse the data in Table 1. Using simulation, we then show in Section 5 that the moment estimates are virtually identical to the maximum likelihood estimates (MLEs) obtained from the random-effects model in a wide variety of situations. We also look at the properties of confidence intervals for the population averages based on asymptotic results. In Section 6, we consider the advantages and disadvantages of using repeated measurements. In the final section, we summarize the main results and discuss some additional issues.

The programs that were used to analyse the data can be obtained from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. A model for repeated measurements with two correlated diagnostic tests

For ease of presentation, we consider a case–control sampling scheme in which the numbers of diseased and non-diseased subjects are fixed, i.e. subjects are selected for the study after the gold standard has been administered. The malaria study is a cohort study in which the numbers of diseased and non-diseased subjects in the study were dependent on the disease prevalence. We can use the analysis derived for the case–control study in a cohort study by conditioning on the number of subjects who are diseased and non-diseased.

To define the notation, we consider only diseased subjects and their corresponding sensitivities. With a simple change of notation, which is described at the end of Section 3, we obtain similar results for non-diseased subjects and their specificities. All probability statements are implicitly conditioned on the disease state. Here, for simplicity, we give results for a study where the number of repeated measurements is the same for all subjects. In Section 7, we describe how the results are easily extended to the more general situation where there may be unequal numbers of repeated measurement on each subject.

If diseased subject i $(i=1,...,n)$ is measured r_k times by diagnostic test k (1 or 2), we use the random variable Y_{ijk} to indicate the result where

$$Y_{ijk}=\begin{cases}1 & \text{if subject } i \text{ tests positively on the } j\text{th repeat of test } k,\\ 0 & \text{otherwise.}\end{cases}$$

Let $\alpha_{ik}=P(Y_{ijk}=1)$ be the sensitivity of the kth test for diseased subject i. Here, the sensitivity α_{ik} is subject specific and we interpret it as the proportion of times that the subject would be correctly diagnosed in a series of many repeats by test k. We assume that the probability of correct diagnosis may vary from subject to subject because of underlying latent covariates that affect the properties of the tests. In many cases, it is unreasonable to assume that sensitivity is constant over all diseased subjects. Some diseased subjects are more difficult to diagnose correctly than others.

To specify the joint distribution for the results of testing the ith subject given the test subject-specific sensitivities α_{i1} and α_{i2}, we assume conditional independence both within and between tests and write

$$P(S_{i1}=s_{i1},S_{i2}=s_{i2}|\alpha_{i1},\alpha_{i2})=\binom{r_1}{s_{i1}}\binom{r_2}{s_{i2}}\alpha_{i1}^{s_{i1}}(1-\alpha_{i1})^{r_1-s_{i1}}\alpha_{i2}^{s_{i2}}(1-\alpha_{i2})^{r_2-s_{i2}} \qquad (1)$$

where $s_{ik}=\Sigma_{j=1}^{r_k}y_{ijk}$ is the number of positive test results for test k on the ith subject and S_{ik} is the corresponding binomial random variable.

Note that with model (1), for each subject, we assume conditional independence between all the measurements made by both measurement systems *given the disease status and the subject test-specific sensitivities*. This model avoids the widespread criticism of the assumption of conditional independence given *only* the diseased status. See for instance Vacek (1983), Fujisawa and Izumi (2000), Pepe (2003), Van Wieringen and de Mast (2008) and de Mast *et al.* (2011) who discussed the conditional independence assumption.

We treat the varying sensitivities as random effects, i.e. we suppose that, for any subject with the disease, the sensitivities, i.e. α_{i1} and α_{i2}, are sampled from a joint density $f(\alpha_1,\alpha_2)$. We discuss some particular choices for $f(\alpha_1,\alpha_2)$ in Section 5. Since we assume that the subject-specific sensitivities are sampled from the same distribution for each subject, in the remainder of this section we simplify the notation by suppressing the i-subscript and provide results for *any* diseased subject. Since the subject-specific sensitivities are not observed, the unconditional joint probability distribution of S_1 and S_2 for a diseased subject is

$$P(S_1=s_1,S_2=s_2)=\binom{r_1}{s_1}\binom{r_2}{s_2}\int\int \alpha_1^{s_1}(1-\alpha_1)^{r_1-s_1}\alpha_2^{s_2}(1-\alpha_2)^{r_2-s_2}f(\alpha_1,\alpha_2)\,\mathrm{d}\alpha_1\,\mathrm{d}\alpha_2. \quad (2)$$

To complete the modelling, we assume that measurements made on different subjects are independent.

We expect the subject-specific sensitivities for the two measurement systems to be correlated when the mechanisms that drive the two tests have common elements. The dependence between repeated measurements on the same subject by the same test is induced by the distribution of the random-effects models. For a diseased (or non-diseased) subject, the correlated random-effects generate a different dependence between two measurements by the same test and two measurements by different tests. If we have two measurements by using the first test, then for any diseased subject $P(Y_{11}=1,Y_{21}=1)=E[\alpha_1^2]$ depending only on the marginal distribution of α_1. If we have two measurements on the same subject, one by each of the tests, then $P(Y_{11}=1,Y_{12}=1)=E[\alpha_1\alpha_2]$, dependent on the joint distribution of α_1 and α_2.

Let $\mu_k=P(Y_{jk}=1)=E[\alpha_k]$, $k=1,2$. The parameters μ_1 and μ_2 represent the population-average sensitivities for the two tests. In the comparison study, the ratio μ_1/μ_2 (or the difference $\mu_1-\mu_2$) is of primary importance. We may also be interested in comparing the variation of the subject-specific sensitivities since a test with less variation in the sensitivities, all else being equal, is preferable. We denote σ_k as the standard deviation of α_k for the kth test and let ρ be the correlation between α_1 and α_2. In some situations, we may also be interested in estimating these secondary parameters. Note also that $\sigma_k^2=\mathrm{var}(\alpha_k)=P(Y_{1k}=1,Y_{2k}=1)-P(Y_{1k}=1)\,P(Y_{2k}=1)$ is a necessarily positive measure of dependence between two measurements by the same test on the same diseased subject. Similarly, $\sigma_{12}=\mathrm{cov}(\alpha_1,\alpha_2)=P(Y_{11}=1,Y_{12}=1)-P(Y_{11}=1)\,P(Y_{12}=1)$ is a measure of dependence between two measurements on the same diseased subject by different tests. The parameters $(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ are determined from the joint density $f(\alpha_1,\alpha_2)$.

3. Estimation

Albert (2007), Albert and Dodd (2008) and Danila *et al.* (2013) found that simple moment estimates were efficient relative to the MLEs from several random-effects models when assessing the properties of a single diagnostic test with repeated measurements on each subject. Here, we extend this approach to the bivariate case. Suppose that there are n diseased subjects in the study. For subject i and test k, we estimate the subject-specific sensitivity by $\hat{\alpha}_{ik}=s_{ik}/r_k$ and the population-average sensitivity by

$$\hat{\mu}_k = \sum_{i=1}^{n} \hat{\alpha}_{ik}/n, \qquad k=1,2. \tag{3}$$

We explore the properties of the estimators corresponding to equation (3) by using the assumption of independence across subjects and by conditioning on the subject-specific sensitivities. Recall that S_{ik} is the random variable corresponding to s_{ik}. Given $(\alpha_{i1}, \alpha_{i2})$, the sensitivities for subject i, S_{ik} is binomial with parameters r_k and α_{ik}. By assumption, S_{i1} and S_{i2} are (conditionally) independent. We have

$$E[\hat{\mu}_k] = \sum_{i=1}^{n} E[\hat{\alpha}_{ik}]/n = \mu_k,$$

so $\hat{\mu}_k$ is unbiased. Also,

$$\begin{aligned} \mathrm{var}(\hat{\mu}_k) &= \frac{\sum_{i=1}^{n} \{E[\mathrm{var}(\hat{\alpha}_{ik}|\alpha_{ik})] + \mathrm{var}(E[\hat{\alpha}_{ik}|\alpha_{ik}])\}}{n^2} \\ &= \frac{\mu_k(1-\mu_k)/r_k + \sigma_k^2(r_k-1)/r_k}{n} \end{aligned} \tag{4}$$

and similarly

$$\mathrm{cov}(\hat{\mu}_1, \hat{\mu}_2) = \mathrm{cov}(\alpha_1, \alpha_2)/n = \sigma_{12}/n. \tag{5}$$

We estimate the standard deviations σ_1 and σ_2, the covariance σ_{12} and the correlation ρ by using equations (4) and (5) as follows. The estimators $\hat{\alpha}_{1k}, \ldots, \hat{\alpha}_{nk}$ are independent and identically distributed and so the sample variance of these estimates is an unbiased estimator of the numerator of equation (4). Substituting $\hat{\mu}_k$ for μ_k and rearranging, we have

$$\hat{\sigma}_k^2 = \frac{r_k}{r_k-1} \left\{ \frac{\sum_{i=1}^{n} (\hat{\alpha}_{ik} - \hat{\mu}_k)^2}{n-1} - \frac{\hat{\mu}_k(1-\hat{\mu}_k)}{r_k} \right\}, \qquad k=1,2. \tag{6}$$

As well, the sample covariance of $(\hat{\alpha}_{i1}, \hat{\alpha}_{i2})$, $i=1,\ldots,n$, is an unbiased estimator of $\sigma_{12} = \mathrm{cov}(\alpha_1, \alpha_2)$ and we have

$$\begin{aligned} \hat{\sigma}_{12} &= \frac{\sum_{i=1}^{n} (\hat{\alpha}_{i1} - \hat{\mu}_1)(\hat{\alpha}_{i2} - \hat{\mu}_2)}{n-1}, \\ \hat{\rho} &= \frac{\sum_{i=1}^{n} (\hat{\alpha}_{i1} - \hat{\mu}_1)(\hat{\alpha}_{i2} - \hat{\mu}_2)}{(n-1)\hat{\sigma}_1\hat{\sigma}_2}. \end{aligned} \tag{7}$$

If $r_k = 1$ for either test ($k=1$ or $k=2$), then equation (4) reduces to the binomial variance and we have no information about σ_k or ρ. We can estimate μ_k and σ_{12} and thus $\mathrm{var}(\hat{\mu}_1)$ and $\mathrm{cov}(\hat{\mu}_1, \hat{\mu}_2)$ by substitution in equation (3) and the first part of equation (5). If we have no repeated measurements for either test, i.e. $r_1 = r_2 = 1$, the model reduces to a multinomial distribution as studied by Nofuentes and del Castillo (2007).

Using the estimates given by equations (4) and (7), we obtain a standard error for the estimate of the difference in sensitivities $\hat{\mu}_1 - \hat{\mu}_2$ or, by using the delta method (Casella and Berger, 2002), a standard error for the ratio $\hat{\mu}_1/\hat{\mu}_2$ or its logarithm. We illustrate the approach by considering the ratios of the average sensitivities on the log-scale, i.e. we estimate $\lambda = \log(\mu_1/\mu_2)$. There are no differences between the population-average sensitivities if $\lambda = 0$. We choose a log-scale for statistical convenience. We can transform the estimates and corresponding confidence intervals to the ratio scale if desired. We have

$$\mathrm{var}(\hat{\lambda}) \approx \frac{\mathrm{var}(\hat{\mu}_1)}{\mu_1^2} + \frac{\mathrm{var}(\hat{\mu}_2)}{\mu_2^2} - 2\frac{\mathrm{cov}(\hat{\mu}_1, \hat{\mu}_2)}{\mu_1\mu_2}. \tag{8}$$

To obtain the approximate standard error, we substitute the estimates (3), (4) and (7) into the square root of approximation (8). We expect these approximations to work well when the number of diseased subjects in the sample is large (see Section 5).

Note that, by a simple change in notation, we have similar results for the specificities. For instance, for the ith non-diseased subject we estimate the test subject specificity as $\hat{\beta}_{ik} = (r_k - s_{ik})/r_k$. We can then use expressions (3), (6) and (7) with $\hat{\alpha}_{ik}$ replaced by $\hat{\beta}_{ik}$ to estimate the population average, variance and correlation for the specificities.

4. Application to comparing two rapid malaria tests

In this section we apply the model that was proposed in Section 2 to the study of Hopkins *et al.* (2007) conducted to compare two rapid malaria tests HRP2 and pLDH. In this study each subject was tested two times: once each by two different raters, by each of the two tests. In our analysis, we ignore any rater effects, treating them simply as repeated measurements. The data of Hopkins *et al.* (2007) are given in Table 1.

Since this application involves both sensitivity and sensitivity, we extend the earlier notation so that β refers to the specificity. Our goal is to compare the average sensitivities $\mu_{\alpha 1}$ and μ_{a2} (denoted by μ_1 and μ_2 in Section 3) and average specificities $\mu_{\beta 1}$ and $\mu_{\beta 2}$ for the two malaria tests.

This was a cohort study in which the number of diseased and non-diseased subjects was not fixed by design. We can use the results from Section 3 by conditioning on the number of diseased subjects: a statistic that is ancillary to the parameters of the two bivariate random-effects distributions. Using equation (3), we obtain the following estimates (denoting HRP2 as the first test and pLDH as the second test):

$$\begin{aligned}\hat{\mu}_{\alpha 1} &= 0.936\,(0.0136),\\ \hat{\mu}_{\beta 1} &= 0.929\,(0.0099),\\ \hat{\mu}_{\alpha 2} &= 0.865\,(0.019),\\ \hat{\mu}_{\beta 2} &= 0.999\,(0.0008).\end{aligned}$$

In parentheses we give approximate standard errors for each estimate derived by substituting the parameter estimates in equation (4). The point estimates are close to those given in Hopkins *et al.* (2007), where they looked at each class of rater separately. With these results, approximate 95% confidence intervals for λ_α and λ_β are 0.079±0.038 and –0.073±0.021 respectively. Transforming to the ratio scale gives approximate 95% confidence intervals of (1.04, 1.12) and (0.91, 0.95) for $\mu_{\alpha 1}/\mu_{\alpha 2}$ and $\mu_{\beta 1}/\mu_{\beta 2}$ respectively. We conclude that the two measurement systems are statistically different. These results support the conclusions that were made in Hopkins *et al.* (2007) that HRP2 had superior sensitivity but inferior specificity when compared with pLDH.

We can also estimate the underlying variability of the subject test-specific sensitivities and specificities and their correlation. We obtain the following point estimates:

$$\begin{aligned}\hat{\sigma}_{\alpha 1} &= 0.22\,(0.027),\\ \hat{\sigma}_{\beta 1} &= 0.24\,(0.018),\\ \hat{\sigma}_{\alpha 2} &= 0.31\,(0.022),\\ \hat{\sigma}_{\beta 2} &= 0.0008\,(0.0004),\\ \hat{\rho}_{\alpha} &= 0.57\,(0.086),\\ \hat{\rho}_{\beta} &= -0.29\,(0.14).\end{aligned}$$

The standard errors (in parentheses) are based on 5000 bootstrap samples where we resampled with replacement from the set of diseased (and non-diseased) subjects.

Other than for the specificity for pLDH (i.e. $\hat{\sigma}_{\beta 2}$), there is considerable variability in the diagnostic rates. There is also substantial correlation between the subject test-specific rates with a positive correlation between diseased subjects and a negative correlation for non-diseased subjects.

As an alternative analysis, we may prefer to compare the positive and negative likelihood ratios (Marshall, 1989) because these ratios correspond to the positive and negative predictive values and may be easier to interpret clinically than sensitivity and specificity. We can use the delta method again to obtain approximate standard errors for the likelihood ratios.

5. Efficiency of the simple estimates

In this section, we explore the efficiency of the simple moment estimates given in Section 3 relative to MLEs. Again, we focus on the sensitivities. To define the likelihood, we must specify the joint density $f(\alpha_1, \alpha_2)$ up to some parameters. Then, the likelihood function is the product of factors given by equation (2) over all n subjects. In the literature, we could find no examples of the use of bivariate correlated random effects for comparing population-average sensitivities and specificities. To assess the efficiency of the moment estimates, we limited consideration to bivariate models in which the marginal distributions of the random effects have been used for assessing a single diagnostic test. We also required that the model allows for possible large (positive or negative) correlation between α_1 and α_2 as seen in the malaria study of Hopkins *et al.* (2007). For this reason, we rejected extensions to a bivariate model with beta–binomial marginal distributions (Danila *et al.*, 2013; Albert and Dodd, 2004), or with Albert and Dodd's (2004) finite mixture marginal distributions. We are left with a bivariate extension to a Gaussian random-effects model based on Qu *et al.* (1996) and Albert and Dodd (2008). We denote this extended model by eGRE. To specify model eGRE, let Z_1 and Z_2 be bivariate Gaussian with means 0, standard deviations 1 and correlation c. Then let $\alpha_1 = \Phi(a_1 + b_1 Z_1)$ and $\alpha_2 = \Phi(a_2 + b_2 Z_2)$ define the subject-specific sensitivities where Φ is the cumulative distribution function of a standard Gaussian random variable. Here

$$\mu_1 = E[\alpha_1] = \Phi\{a_1/\surd(1 + b_1^2)\} \qquad (9)$$

with a similar expression for μ_2 (Qu *et al.*, 1996). There are no simple formulae to express the standard deviations σ_1 and σ_2 or the correlation ρ in terms of a_1, a_2, b_1, b_2 and c but, if desired, they can be determined through numerical integration. With this bivariate extension, the correlation between α_1 and α_2 is flexible and the marginal distributions follow the proposal of Qu *et al.* (1996). For the malaria study from Section 4 we can assess model fit. Table 1 gives

Table 2. Factor levels for simulations 1 and 2

Parameter	*Results for simulation 1 to compare moment estimates and MLEs (768 combinations)*	*Results for simulation 2 to assess moment estimates (2592 combinations)*
Number of subjects	100, 400	100, 200, 400
$r\ (=r_1=r_2)$	2, 3	2, 3
μ_{α_1}	0.7, 0.9	0.7, 0.8, 0.9
μ_{α_2}	0.7, 0.9	0.7, 0.8, 0.9
σ_{α_1}	0.02, 0.1, 0.2, 0.3	0.02, 0.1, 0.2, 0.3
σ_{α_2}	0.02, 0.1, 0.2, 0.3	0.02, 0.1, 0.2, 0.3
ρ	−0.5, 0, 0.5	−0.5, 0, 0.5

the expected cell counts by using model eGRE and the parameter estimates presented in Section 4. We see that there is very good agreement—the model fits the data very well. Of course, there are many other models, which are not considered here, that may fit the data equally well.

We considered two simulation scenarios. For both, we generated data from model eGRE and restricted our attention to sensitivities. In the first simulation, we compare the simple moment estimates that were given in Section 3 with the MLEs from model eGRE. We used a full factorial arrangement of design and parameter values with levels shown in Table 2. We limited the number of combinations to 768 because finding the MLEs of model eGRE is time consuming. The number of diseased subjects is 100 or 400, which are comparable with the number of positive subjects in the malaria study. For each of the 768 combinations, we first determined numerically the corresponding values of a_1, a_2, b_1, b_2 and c. Then, we generated a single set of sample data from model eGRE and calculated the estimates from equations (3), (6) and (7) as well as the MLEs for model eGRE (by using numerical optimization). This simulation extended the results from Albert and Dodd (2008) and Danila *et al.* (2013) who showed that the moment estimates are efficient in the univariate (single-test) case. In cases where μ_1, μ_2, σ_1 and σ_2 are all simultaneously large it is not possible to generate data from model eGRE with a correlation of $\rho=-0.5$. In these cases, in the simulation, we increased the desired value of ρ in increments of 0.1 until we could find values of a_1, a_2, b_1, b_2 and c that matched the desired parameter values. In Fig. 1, we plot the simple moment estimates against the corresponding MLEs. We see that the simple moment estimates of μ_1 and μ_2, i.e. the population-average sensitivities, are virtually identical to the MLEs. This implies that they can be used interchangeably. The relationships between the two estimates of the other parameters σ_1 and σ_2 are somewhat noisier, however. We do not show the estimates for ρ because when the estimate for either σ_1 or σ_2 equals 0 it is not defined. Note that the estimate from equation (6) for σ_k^2 can be negative. Because of this problem, the moment estimates and MLEs for σ_k did not agree that well. However, the estimates of $\text{cov}(\alpha_1, \alpha_2)$ from the two approaches are similar. These simulation results suggest that the moment estimates from Section 3 (at least for the sensitivity, specificity and covariance) are efficient relative to the MLEs based on model eGRE. The moment estimates (3) for the parameters of primary interest are simple to calculate and have an easy-to-calculate measure of precision. In contrast, the MLEs and standard errors based on observed and expected information are difficult to calculate and we have not investigated in detail for different parameter values how large the sample size must be for the standard errors based on asymptotic results to be approximately correct.

In the second simulation, we considered only the moment estimates to see how well they estimate the true parameter values. For each of the 2592 combinations of the parameter values

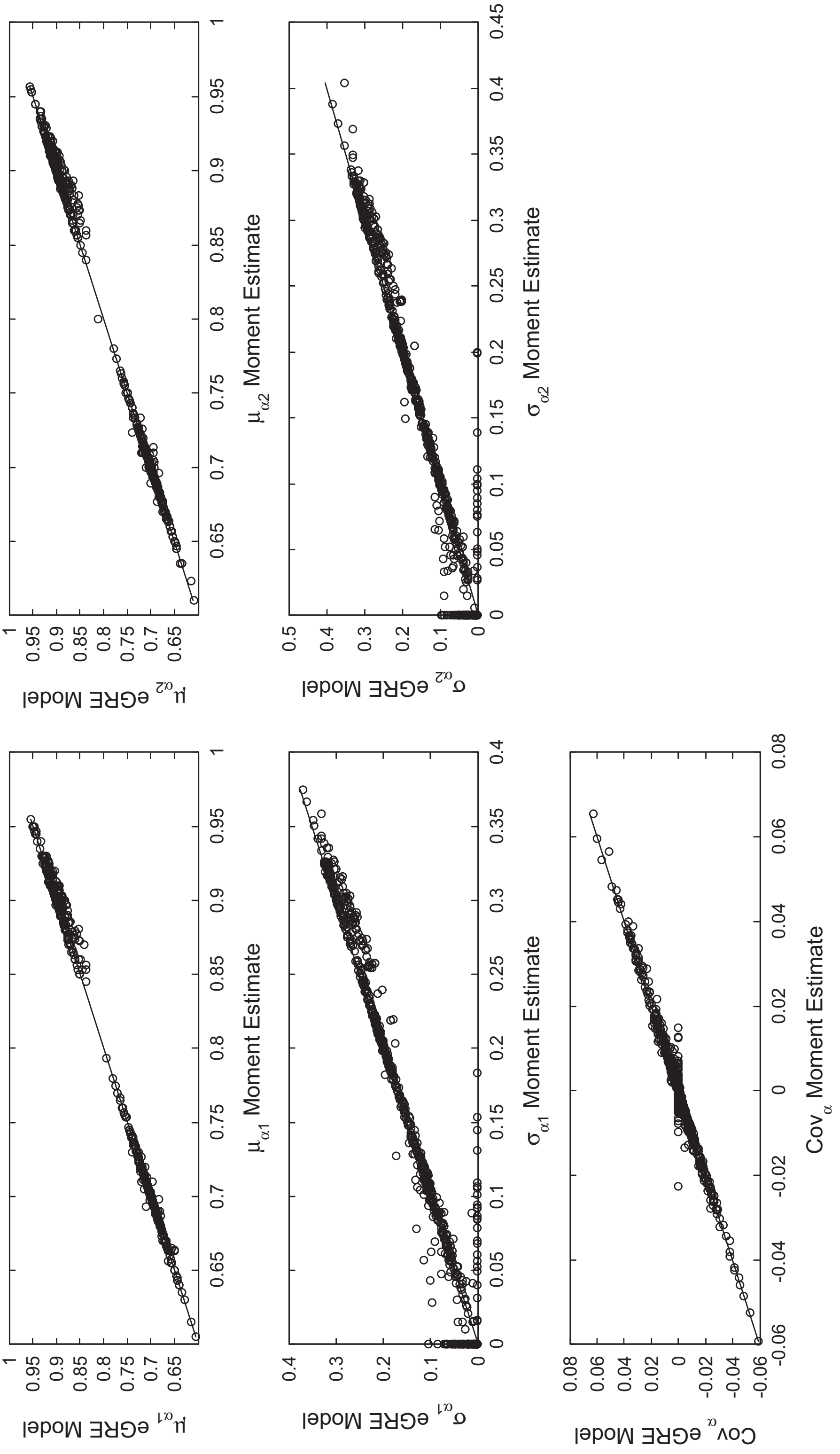

Fig. 1. Comparison of moment estimates from Section 3 and MLEs from model eGRE

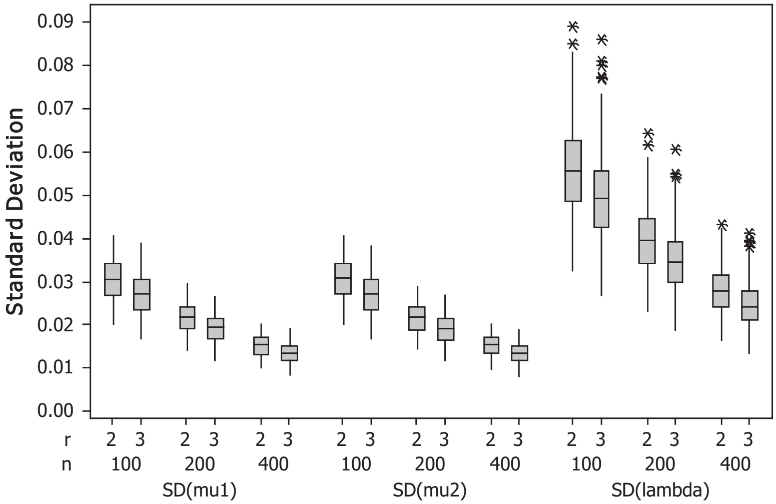

Fig. 2. Moment estimate standard deviations for simulation 2

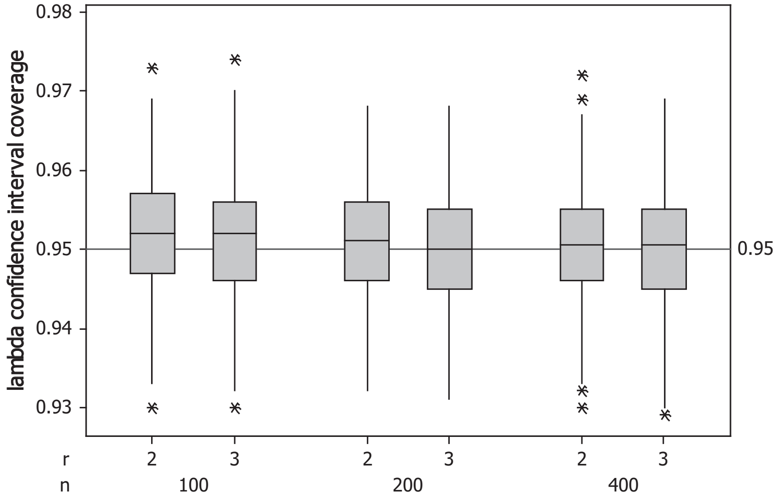

Fig. 3. Boxplots of coverage for the approximate 95% confidence interval for λ_α from simulation 2

in Table 2, we generated 1000 samples from the corresponding eGRE model and calculated the moment estimates for μ_k, σ_k, $k=1,2$, ρ, λ_α (as defined in Section 4) and an approximate 95% confidence interval for λ_α by using the standard error derived from equation (8). Just as in the first simulation study, in cases where μ_1, μ_2, σ_1 and σ_2 are all simultaneously large when the desired $\rho=-0.5$, we increased the desired value of ρ in increments of 0.1 until we could find values of a_1, a_2, b_1, b_2 and c that matched the desired parameter values. This happened in about 5% of the runs. The simulation results suggest that the moment estimates are unbiased for μ_1, μ_2, σ_{12}, ρ and λ_α, whereas the biases for estimating σ_1 and σ_2 can be large with worse results when the true σ_1- and σ_2-values are small. This is because of our decision to set negative estimates to 0. Fig. 2 shows the standard deviations of the estimators corresponding to the three parameters μ_1, μ_2 and λ_α stratified by sample size and number of repeated measurements from 1000 runs for each of the combinations. We see that the mean sensitivities and log-ratio of the mean sensitivities are well estimated across all combinations of parameter values considered. In addition, as expected, the standard deviations decrease with larger sample sizes and more repeated measurements.

Fig. 3 illustrates the performance of the approximate 95% confidence interval for λ_α. We see that coverage of the confidence interval (i.e. the proportion of confidence intervals created in each of the 1000 runs that include the true parameter value) is very close to the desired 0.95 across all parameter combinations.

On the basis of these simulation results, we recommend using the simple moment estimates. The moment estimate for the correlation is poorly behaved unless we use very large samples sizes and many repeated measurements. However, the approximate confidence intervals for the comparison based on the logarithm of the mean sensitivities (or specificities) are well behaved even for small sample sizes and few repeated measurements.

6. Advantages of using repeated measurements

Here we want to discuss the advantages and disadvantages of repeating the tests on each subject. The disadvantages seem to be ethical or logistical. In many cases, the tests may be invasive and we cannot then repeat each test. However, from a statistical and planning perspective, there are advantages to repeating the tests on each subject. Again we consider only the sensitivities with similar conclusions about the specificities. Qualitatively, it may be difficult to acquire subjects for the study, especially if the gold standard test is expensive or invasive. If repetition is possible, we can increase precision of the comparisons by repeating the tests. Suppose that each of the n subjects is measured r times with each test (i.e. $r_1 = r_2$). With a little algebra using equations (4) and (5), we obtain

$$\begin{aligned}\operatorname{var}(\hat{\mu}_1 - \hat{\mu}_2) &= \frac{1}{n}\left\{\frac{\mu_1(1-\mu_1)}{r} + \sigma_1^2\frac{r-1}{r} + \frac{\mu_2(1-\mu_2)}{r} + \sigma_2^2\frac{r-1}{r} - 2\rho\sigma_1\sigma_2\right\} \\ &= \frac{\text{constant} + r\operatorname{var}(\alpha_1 - \alpha_2)}{nr}.\end{aligned}$$

If we hold n, the number of subjects, fixed, increasing r reduces the variability of the estimate. However, if we hold $2nr$, the total number of tests, fixed, increasing r increases the variance of the estimate of $\mu_1 - \mu_2$ regardless of the correlation between α_1 and α_2. As expected, a repeated measurement is less valuable in terms of precision than is an extra subject. For $\operatorname{var}(\hat{\lambda})$, we obtain a similar expression (the constant is changed and α_i is replaced by α_i/μ_i) and the same conclusions. If we do not replicate, we cannot estimate the secondary parameters σ_1, σ_2 or ρ and so we would have no idea about the variation of the sensitivities within the study population. We temper this statement recognizing that we require very large samples to estimate these parameters well.

In the above discussion, we assumed that $r_1 = r_2$. In other cases, it is not clear how $\operatorname{var}(\hat{\mu}_1 - \hat{\mu}_2)$ or $\operatorname{var}(\hat{\lambda})$ depend on n, r_1 and r_2 since these measures of precision also depend on the underlying parameter values. For example, suppose that $\mu_{\alpha 1} = 0.9$, $\mu_{\alpha 2} = 0.8$, $\sigma_{\alpha 1} = 0.05$, $\sigma_{\alpha 2} = 0.075$ and $\rho_\alpha = 0.5$. Then, for different combinations of r_1 and r_2, using approximation (8), we construct Table 3 to give the minimum number of diseased subjects n so that the standard deviation of $\hat{\lambda}_\alpha$ is as large as possible but less than or equal to 0.05. Because n is large, all of the displayed plans have $\operatorname{Stdev}(\hat{\lambda}_\alpha)$ very close to 0.05. We see, for example, that if we measured 95 subjects once with the first test and twice with the second ($r_1 = 1$ and $r_2 = 2$) we obtain a substantial reduction in the number of subjects required compared with the case $r_1 = r_2 = 1$ (no repeated test measurements). Furthermore, the total number of test measurements is almost the same.

Table 3. Minimum number of diseased subjects (and total number of test measurements in parentheses) to achieve StDev($\hat{\lambda}_\alpha$) $\leqslant$ 0.05 under various combinations of r_1 and r_2 (replications by tests 1 and 2 respectively) when $\mu_{\alpha 1}$ = 0.9, $\mu_{\alpha 2}$ = 0.8, $\sigma_{\alpha 1}$ = 0.05, $\sigma_{\alpha 2}$ = 0.075 and ρ_α = 0.5

r_2	*Results for the following values of r_1:*			
	1	*2*	*3*	*4*
1	143 (286)	121 (363)	114 (456)	110 (550)
2	95 (285)	73 (292)	66 (330)	62 (372)
3	79 (316)	57 (285)	50 (300)	46 (322)
4	70 (350)	49 (294)	42 (294)	38 (304)

Table 4. Optimal cost plan that achieves StDev ($\hat{\lambda}_\alpha \leqslant$ 0.05) for values of *c* when $\mu_{\alpha 1}$ = 0.9, $\mu_{\alpha 2}$ = 0.8, $\sigma_{\alpha 1}$ = 0.05, $\sigma_{\alpha 2}$ = 0.075 and ρ_α = 0.5

c	*Results for the optimal plan*				*Cost for $r_1 = r_2 = 2$*	*Cost for $r_1 = r_2 = 1$*
	n	r_1	r_2	*Cost*		
1	3	4	5	330	365	429
2	22	6	8	352	438	572
3	22	6	8	374	511	715

There is a substantial reduction in the number of required diseased subjects that comes from using multiple measurements by the second test in this example. Remember that we assess each selected subject with the gold standard which may be expensive and invasive. If making repeated measurements is less expensive than selecting additional subjects, using repeated tests may substantially reduce the overall cost of the study. In the example given in Table 3, we can meet the precision criterion by using only 57 subjects with $r_1 = 2$ and $r_2 = 3$ without an increase in the required total number of test measurements.

The benefit of repeated measurement is more pronounced if we consider a model in which the cost of a gold standard measurement (including subject recruitment) is greater than that of measurement by either of the two tests. Suppose that we scale the cost so that a repeated measurement by either diagnostic test costs 1 monetary unit and a single measurement by the gold standard costs c monetary units. We expect $c > 1$. Then, the cost of measurement of the proposed study is $n(c + r_1 + r_2)$, where n is the number of subjects and r_1 and r_2 are the numbers of repeated measurements by tests 1 and 2. For simplicity, we assume that measurements by the two diagnostic tests cost the same. We now consider plans that minimize this cost for various values of c, subject to a constraint on the precision of the comparison. Table 4 provides results for the parameter values that were considered in Table 3. Again, we look only at the required number of diseased subjects (n in Table 4) for plans with StDev($\hat{\lambda}_\alpha$) as close as possible but less than or equal to 0.05.

Since the optimal number of repeated measurements rises rapidly, we also include the more plausible case with $r_1 = r_2 = 2$ to demonstrate the value of using repeated measurements.

In planning a comparison study when repeated measurements are possible, we recommend setting precision requirements and then investigating var($\hat{\lambda}$) or var($\hat{\mu}_1 - \hat{\mu}_2$) (and the corresponding quantities for specificities) for various values of n, r_1 and r_2 and a range of plausible values of the underlying parameters. The approximate variance var($\hat{\lambda}_\alpha$), given by approximation (8), applies directly to the case–control study where the numbers of diseased and non-diseased subjects are prespecified. If we instead select a random sample of m subjects (i.e. use a cohort study) then we replace n by $m\theta$ in expressions (4) and (5) where θ is the prevalence of disease in the study population.

7. Summary and discussion

We examined the analysis and planning of a study to compare two diagnostic tests when a gold standard is available and it is possible to make repeated measurements by each test on individual subjects. We can apply the results when we select the subjects for the study at random from the target population (i.e. use a cohort study) or when we randomly select a fixed number of diseased and non-diseased subjects from the population premeasured with the gold standard (i.e. use a case–control study). We propose a bivariate random-effects model that incorporates varying subject test-specific sensitivities (and separately specificities). For each diseased or non-diseased subject, the models include a correlation between the sensitivities or specificities for each test. We showed that simple closed form moment estimates of the average sensitivity and specificity are virtually identical to the MLEs for a particular random-effects model based on the bivariate normal distribution.

We give simple expressions that can be used to derive standard errors of the moment estimates, regardless of the underlying random-effects models. These lead to estimates and approximate standard errors for the ratio of the average sensitivities (and specificities) between the two tests under study. For the estimates of the secondary parameters σ_1, σ_2 and ρ, we suggest using bootstrapping to obtain approximate standard errors, as performed when comparing two rapid malaria tests in Section 4.

We use these results to aid in planning a comparison study in terms of the number of subjects and number of repeated measurements per subject for each test. Depending on the parameter values, we may see a large reduction in the number of subjects required and the total cost if we measure each subject more than once with each diagnostic test. We can achieve this reduction without increasing the total number of test measurements.

We considered study plans in which the number of repeated measurements for a particular test was the same for each subject. If this number varies from subject to subject, we can use the moment estimates of the average sensitivities (or specificities), but we need to adjust the estimates of the variances (4) and covariance (5). In the case where we denote the number of measurements by test k on subject i as r_{ik} the numerator of equation (4) becomes

$$\frac{\mu_k(1-\mu_k)\sum_{i=1}^{n} 1/r_{ik} + \sigma_k^2 \sum_{i=1}^{n} (r_{ik}-1)/r_{ik}}{n}.$$

This result allows derivation of a result similar to equation (6) when there are unequal measurements per subject for either test.

We have presented results for comparing two diagnostic tests. We can extend the moment estimates given in equations (3), (6) and (7) to the comparison of $k \geqslant 3$ tests. The joint distribution of the estimates of the average sensitivities is approximately multivariate normal with variances and covariances given by equations (4) and (5). We can use this approximation to construct a test of the hypothesis that there are no differences between the population-average sensitivities for the k tests.

We examined studies where the true disease status of every subject is verified through the use of a gold standard. If there is partial verification (Albert and Dodd, 2008), then the moment estimates are no longer available and we must resort to selecting a specific model and using maximum likelihood estimation. If no gold standard system is available, then we must use a cohort study and maximum likelihood estimation for the latent class model generated by the underlying random-effects distribution. We have not investigated the extended Gaussian random effects model eGRE in these cases. However, Albert and Dodd (2004) showed that, with latent class models, we need to exercise extreme caution since the MLEs of the primary parameters can be severely biased if we fit the wrong model. Unfortunately, in such cases, there is little information to assess any assumed model.

In industry, manufacturers inspect product with go–no-go gauges (i.e. gauges that check a part against its allowed tolerances) and functional gold standard tests that provide a pass or fail determination. The key goal is to protect the customer from receiving non-conforming product. The results in this paper can be directly applied in the manufacturing context.

Acknowledgements

The authors thank Heidi Hopkins for sharing the detailed data from the study to compare malaria tests reported in Hopkins *et al.* (2007). We also thank the Joint Editor, Associate Editor and referees for numerous positive suggestions. This research was supported, in part, by research grants (105240 and 402474) from the Natural Sciences and Engineering Research Council of Canada.

References

Albert, P. S. (2007) Imputation approaches for estimating diagnostic accuracy for multiple tests from partially verified designs. *Biometrics*, **63**, 947–957.

Albert, P. S. and Dodd, L. E. (2004) A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, **60**, 427–435.

Albert, P. S. and Dodd, L. E. (2008) On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *J. Am. Statisti. Ass.*, **103**, 61–73.

Baker, S. G., Freedman, L. D. and Parmar, M. K. B. (1991) Using replicate observations in observer agreement studies with binary assessments. *Biometrics*, **47**, 1327–1338.

Biggerstaff, B. J. (2000) Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statist. Med.*, **19**, 649–663.

Casella, G. and Berger, R. L. (2002) *Statistical Inference*, 2nd edn. Andover: Cengage Learning.

Danila, O., Steiner, S. H. and MacKay, R. J. (2013) Assessing a binary measurement system with varying misclassification rates when a gold standard is available. *Technometrics*, **55**, 335–345.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.

Fujisawa, H. and Izumi, S. (2000) Inference about the misclassification probabilities from repeated binary responses. *Biometrics*, **56**, 706–711.

Hopkins, H., Kambale, W., Kamya, M., Staedke, S., Dorsey, G. and Rosenthal, P. (2007) Comparison of HRP2- and pLDh-based rapid diagnostic tests for malaria with longitudinal follow-up in Kampala. *Ug. Am. J. Trop. Med. Hyg.*, **76**, 1092–1097.

Marshall, R. J. (1989) The predictive value of simple rules for combining two diagnostic tests. *Biometrics*, **45**, 1213–1222.

de Mast, J., Erdmann, T. P. and Van Wieringen, W. (2011) Measurement system analysis for binary inspection: continuous versus dichotomous measurands. *J. Qual. Technol.*, **43**, 99–112.

Nofuentes, J. A. and Del Castillo, D. (2007) Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. *Statist. Med.*, **26**, 4179–4201.

Pepe, M. S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*, 1st edn. New York: Oxford University Press.

Qu, Y., Tan, M. and Kutner, M. H. (1996) Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, **52**, 797–810.

Vacek, P. (1983) The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, **41**, 959–968.

Van Wieringen, W. N. and de Mast, J. (2008) Measurement system analysis for binary data. *Technometrics*, **50**, 468–478.

Zhou, X. H., Obuchowski, N. A. and McClish, D. K. (2011) *Statistical Methods in Diagnostic Medicine*. New York: Wiley.