

## Joint Estimation of Diagnostic Accuracy Measures for Paired Organs – Application in Ophthalmology

Alexander R. de Leon<sup>1,\*</sup>, Andrea Soo<sup>1</sup>, Daniel C. Bonzo<sup>2</sup> and Christopher J. Rudnisky<sup>3</sup>

<sup>1</sup> Department of Mathematics and Statistics, University of Calgary, Calgary, Alta., T2N 1N4, Canada

<sup>2</sup> Biometrics and Data Management Unit, Xenoport, Santa Clara, CA 95051, USA

<sup>3</sup> Department of Ophthalmology, University of Alberta, Edmonton, Alta., T6G 2G1, Canada

Received 12 June 2008, revised 5 July 2009, accepted 6 July 2009

Diagnostic studies in ophthalmology frequently involve binocular data where pairs of eyes are evaluated, through some diagnostic procedure, for the presence of certain diseases or pathologies. The simplest approach of estimating measures of diagnostic accuracy, such as sensitivity and specificity, treats eyes as independent, consequently yielding incorrect estimates, especially of the standard errors. Approaches that account for the inter-eye correlation include regression methods using generalized estimating equations and likelihood techniques based on various correlated binomial models. The paper proposes a simple alternative statistical methodology of jointly estimating measures of diagnostic accuracy for binocular tests based on a flexible model for correlated binary data. Moments' estimation of model parameters is outlined and asymptotic inference is discussed. The resulting estimates are straightforward and easy to obtain, requiring no special statistical software but only elementary calculations. Results of simulations indicate that large-sample and bootstrap confidence intervals based on the estimates have relatively good coverage properties when the model is correctly specified. The computation of the estimates and their standard errors are illustrated with data from a study on diabetic retinopathy.

*Key words:* Common correlation model; Correlated binary data; Coverage probability; Moments estimation; Predictive values; Sensitivity; Specificity.

Supporting Information for this article is available from the author or on the WWW under <http://dx.doi.org/10.1002/bimj.200800123>.

### 1 Introduction

The motivation for this paper is a study conducted in Alberta, Canada, on the use of high-resolution stereoscopic digital photography for evaluating diabetic patients at a distance for treatable diabetic retinopathy (Rudnisky *et al.*, 2002). A teleophthalmology system allowing for distance evaluation of retinopathy-related pathologies based on digital images of diabetic patients' eyes is a potentially cost-effective alternative to clinical examination in countries like Canada, where distances are great and the cost of travel is high. The diagnostic procedure involves digital images of patients' eyes that are evaluated by a trained reader for certain retinopathy-related pathologies. The purpose of the study was to determine whether diabetic retinopathy can be identified with high-resolution stereoscopic digital photography and whether this identification correlates well with the accepted gold standard of clinical examination.

\* Correspondence author: e-mail: [adeleon@math.ucalgary.ca](mailto:adeleon@math.ucalgary.ca), Phone: +1-403-220-6782, Fax: +1-403-282-5150

The accuracy of a medical test for diagnosing the presence or absence of a disease can be described by several measures, the most common of which are given by the test's sensitivity and specificity with respect to the true disease status as determined by a traditionally used and accepted test regarded as a "gold standard." Sensitivity is the probability that the new test indicates presence of the disease when the gold standard indicates that it is present while specificity is the probability that the new test indicates absence of the disease when the gold standard indicates that it is absent. Other frequently used measures of diagnostic accuracy are the so-called post-test probabilities given by the test's positive predictive and negative predictive values. The former is defined as the probability of presence of disease given a positive test result while the latter is the probability of absence of disease given a negative test result (see, *e.g.* Zhou, Obuchowski, and McClish, 2002, pp. 44–45). Positive and negative predictive values describe how well a test predicts a patient's disease status, while sensitivity and specificity describe how well the test discriminates between positive disease status and negative disease status.

Because digital images of both left and right eyes of patients are evaluated, the binocular structure of the data impacts on the analysis, as an eye tends to have a greater correspondence with the fellow eye than with other eyes. While it is possible to estimate a test's diagnostic accuracy on an eye-specific basis, thereby effectively ignoring the inter-eye correlation, incorrect inferences are likely to result from underestimated standard errors (Glynn and Rosner, 1992). Methods that account for this correlation are thus needed. Approaches to handling this problem include the generalized estimating equations (GEE) approach of Smith and Hadgu (1992) (see also Sternberg and Hadgu, 2001; Leisenring, Pepe, and Longton, 1997) and those based on likelihood methods studied by Hujuel, Moulton, and Loesche (1990); Rosner (1989), among others (see also Sutradhar and Das, 1997; Lefkopoulou, Moore, and Ryan, 1989). Recent references on correlated ophthalmologic data include de Leon *et al.* (2007); Leite and Nicolosi (1998). While these approaches are able to accommodate covariates and a variety of dependence structures, they are regression-based and computationally much more demanding.

In this paper, we propose simple easy-to-calculate estimates of measures of diagnostic accuracy and their standard errors using binocular binary data. The method arises from a generalization introduced by Shoukri and Donner (2007) of the common correlation model (CCM) for correlated binary data (Mak, 1988). This general formulation includes the well-known beta-binomial model as well as the correlated binomial models of Haseman and Kupper (1978). We discuss the model in Section 2.

The rest of the paper is organized as follows. Estimation for the model *via* the method of moments as well as of various measures of diagnostic accuracy is discussed in Section 2. Section 3 presents simulation results on the coverage probabilities of large-sample and bootstrap confidence intervals (CIs) for the various measures of diagnostic accuracy considered. The methodology is illustrated in Section 4 on data from the diabetic retinopathy study described earlier. We conclude with a brief discussion in Section 5.

## 2 Binocular Model

Define  $Y_{i1L}$  and  $Y_{i1R}$  as 1 if the reader indicates the presence of the disease in the left and right eyes, respectively, of patient  $i = 1, \dots, N$ , and 0, otherwise. Furthermore, let  $Y_{i2L}$  and  $Y_{i2R}$  denote the true disease status (1 if positive, 0 if negative) of the left and right eyes, respectively, of patient  $i$  as determined by the gold standard. In what follows, we use random effects to flexibly model the joint distribution of  $Y_{i1L}$ ,  $Y_{i1R}$ ,  $Y_{i2L}$ , and  $Y_{i2R}$  and to capture the key features of the correlations between them.

For  $j = 1, 2$ , let the random effects  $P_j = P(Y_{ijL} = 1|P_j) = P(Y_{ijR} = 1|P_j)$  be the common conditional probability of a positive result, and let the conditional distribution of  $(Y_{ijL}, Y_{ijR})^\top$  given random effect  $P_j$ , be a CCM with intra-pair correlation  $\kappa_j$ . Assuming  $(Y_{i1L}, Y_{i1R})^\top$  and  $(Y_{i2L}, Y_{i2R})^\top$

are conditionally independent given the random effects, the (unconditional) joint probability  $P_{\ell_1 r_1 \ell_2 r_2} = P(Y_{i1L} = \ell_1, Y_{i1R} = r_1, Y_{i2L} = \ell_2, Y_{i2R} = r_2)$  is then obtained as

$$\begin{aligned}
 P_{\ell_1 r_1 \ell_2 r_2} &= \int_0^1 \int_0^1 P(Y_{i1L} = \ell_1, Y_{i1R} = r_1 | p_1) P(Y_{i2L} = \ell_2, Y_{i2R} = r_2 | p_2) f(p_1, p_2) dp_1 dp_2, \\
 &= \int_0^1 \int_0^1 P(Y_{i1L} = \ell_1, Y_{i1R} = r_1 | p_1) P(Y_{i2L} = \ell_2, Y_{i2R} = r_2 | p_2) \\
 &\quad \times f_1(p_1) f_2(p_2) \left\{ 1 + \frac{(p_1 - \pi_1)(p_2 - \pi_2)}{\sqrt{\text{var}(P_1)\text{var}(P_2)}} \right\} dp_1 dp_2,
 \end{aligned}
 \tag{1}$$

where we used a canonical representation for the joint density  $f(\cdot, \cdot)$  of  $P_1$  and  $P_2$  (Mardia, 1970). Here,  $f_j(\cdot)$  is the density of  $P_j$ ,  $\pi_1 = P(Y_{i1L} = 1) = P(Y_{i1R} = 1)$  is the probability of a positive diagnosis by the reader, and  $\pi_2 = P(Y_{i2L} = 1) = P(Y_{i2R} = 1)$  is the prevalence of the disease. Taking  $P_j \sim \text{beta}(\alpha_j = (1 - \rho)\pi_j/\rho, \beta_j = (1 - \rho)(1 - \pi_j)/\rho)$ , with  $\rho$  as a ‘correlation’ parameter (see discussion below), Shoukri and Donner (2007) showed that (1) reduces to nine distinct probabilities, which can be written in terms of  $\kappa_1, \kappa_2$ , and the non-central product moments

$$\psi(m_1, m_2) = E(P_1^{m_1} P_2^{m_2}) = E(P_1^{m_1}) E(P_2^{m_2}) \left\{ 1 + \frac{(\gamma_1^{(m_1)} - \pi_1)(\gamma_2^{(m_2)} - \pi_2)}{\rho\tau} \right\},$$

where  $\tau^2 = \pi_1\pi_2(1 - \pi_1)(1 - \pi_2)$ ,

$$E(P_j^{m_j}) = \frac{\Gamma(\alpha_j + m_j)\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\alpha_j + \beta_j + m_j)} \quad \text{and} \quad \gamma_j^{(m_j)} = \frac{(1 - \rho)\pi_j + \rho m_j}{(1 - \rho)\pi_j + (1 - \rho)(1 - \pi_j) + \rho m_j},$$

and  $\Gamma(\cdot)$  is the  $\gamma$  function.

The unconditional distribution of  $(Y_{ijL}, Y_{ijR})^\top$  is a CCM with intra-pair correlation  $\rho_j = \rho + \kappa_j(1 - \rho)$  and common probability  $\pi_j$  of a positive result. Observe that  $\kappa_1 = (\rho_1 - \rho)/(1 - \rho)$  has a  $\kappa$ -like form (Fleiss, Levin, and Paik, 2003, p. 598), so that  $\kappa_1$  can be interpreted as a measure of agreement between the reader’s left- and right-eye diagnoses; similarly,  $\kappa_2$  can be viewed as representing the agreement between the disease status of fellow eyes. Note that  $\rho_1 = \rho$  if and only if  $\kappa_1 = 0$ , which indicates that the reader’s diagnoses are conditionally independent given  $P_1$ , and the correlation  $\rho_1$  between the diagnoses can be mainly attributed to the overall correlation  $\rho$ . Similarly,  $\kappa_2 = 0$  implies conditional independence of disease status of fellow eyes given  $P_2$ , and suggests that the main source of association is  $\rho$ .

It is easy to see that  $\rho$  is the correlation between the number  $Y_{i1\bullet} = Y_{i1L} + Y_{i1R}$  of diagnosis-positive eyes and the number  $Y_{i2\bullet} = Y_{i2L} + Y_{i2R}$  of status-positive eyes. Thus, the correlation  $\rho$  provides a global measure of association between the diagnoses and the disease status of the eyes, albeit aggregated over disease status and diagnoses of fellow eyes. Shoukri and Donner (2009) recently showed that the overall correlation  $\rho$  must satisfy

$$\frac{1}{\kappa_1\kappa_2} \max \left\{ -\frac{1}{\pi_1\pi_2}, -\frac{1}{(1 - \pi_1)(1 - \pi_2)} \right\} \leq \rho \leq \frac{1}{\kappa_1\kappa_2} \min \left\{ -\frac{1}{\pi_1(1 - \pi_2)}, -\frac{1}{(1 - \pi_1)\pi_2} \right\}$$

to ensure that the joint probabilities  $P_{\ell_1 r_1 \ell_2 r_2}$  determined by (1) define a proper probability distribution.

A strength of model (1) is that the joint distribution of the binocular data  $(Y_{i1L}, Y_{i1R})^\top$  and  $(Y_{i2L}, Y_{i2R})^\top$  is completely determined by the marginal densities of  $P_1$  and  $P_2$ . By choosing beta densities as marginal distributions for  $P_1$  and  $P_2$ , we are able to flexibly model a variety of uncertainties regarding the distributions of  $P_1$  and  $P_2$  (Johnson and Kotz, 1970, Chapter 25), and hence of the binocular diagnostic data. In addition, model (1) yields convenient marginal distributions in that  $(Y_{i1L}, Y_{i1R})^\top$  and  $(Y_{i2L}, Y_{i2R})^\top$  are both modeled by CCMs. In this sense, model (1) can be viewed as an extension of CCM to paired binocular binary data. As shown in Section 2.1,

model (1) lends itself to straightforward non-iterative estimation of its parameters by the method of moments.

A potential shortcoming of the model involves the assumption  $P_{1010} = P_{0101} = P_{1001} = P_{0110}$ , which suggests that perfect agreement between diagnosis and status for exactly one eye is equally likely as perfect disagreement. As pointed out by a referee, this may not hold in cases concerning chronic diseases. We note that this equivalence arose from the assumption of exchangeability of fellow eyes, and as such, can possibly be remedied by adopting a richer family of densities for  $P_1$  and  $P_2$  in model (1); generalized versions of the beta distribution (Nadarajah and Kotz, 2007) involving additional parameters for added flexibility are possible choices. In any case, a goodness-of-fit test can be used to validate this assumption in practice. We investigate how violation of this assumption impacts the analysis in Section 3.

Model (1) is most appropriate for analyzing binocular binary data in ophthalmology like those described earlier, as it is able to meaningfully delineate the intra-pair correlations in the binocular data into  $\rho_1$ , the correlation between the reader's left- and right-eye diagnoses, and  $\rho_2$ , the correlation between the disease status of fellow eyes. It provides a useful computationally simple non-iterative alternative to commonly used regression-based methods.

## 2.1 Moments estimation

Suppose  $n_{xy}$  patients having  $y = 0, 1, 2$  eyes with positive disease status have  $x = 0, 1, 2$  eyes diagnosed as positive by the reader, with  $\sum_{x=0}^2 \sum_{y=0}^2 n_{xy} = N$ . Let  $\boldsymbol{\theta} = (\pi_1, \pi_2, \rho_1, \rho_2, \rho)^\top$  and denoting the probabilities  $p_{xy} = P(Y_{i1\bullet} = x, Y_{i2\bullet} = y)$ , we have  $p_{00} = P_{0000}$ ,  $p_{10} = P_{1000} + P_{0100}$ ,  $p_{20} = P_{1100}$ ,  $p_{01} = P_{0010} + P_{0001}$ ,  $p_{11} = P_{1010} + P_{1001} + P_{0110} + P_{0101}$ ,  $p_{21} = P_{1110} + P_{1101}$ ,  $p_{02} = P_{0011}$ ,  $p_{12} = P_{0111} + P_{1011}$ , and  $p_{22} = P_{1111}$ . We thus get the moments estimates  $\hat{p}_{00} = n_{00}/N, \dots, \hat{p}_{22} = n_{22}/N$ . Noting that

$$\begin{aligned}\pi_1 &= \theta_1 = \frac{1}{2}\{p_{01} + p_{11} + p_{21} + 2(p_{02} + p_{12} + p_{22})\}, \\ \pi_2 &= \theta_2 = \frac{1}{2}\{p_{10} + p_{11} + p_{12} + 2(p_{20} + p_{21} + p_{22})\}, \\ \rho_1 &= \theta_3 = 1 - \frac{p_{01} + p_{11} + p_{21}}{2\theta_1(1 - \theta_1)}, \\ \rho_2 &= \theta_4 = 1 - \frac{p_{10} + p_{11} + p_{12}}{2\theta_2(1 - \theta_2)}, \\ \rho &= \theta_5 = \frac{p_{11} + 2p_{21} + 2p_{12} + 4p_{22}}{4\sqrt{\theta_1\theta_2(1 - \theta_1)(1 - \theta_2)}} - \frac{\theta_1\theta_2}{\sqrt{\theta_1\theta_2(1 - \theta_1)(1 - \theta_2)}},\end{aligned}$$

*i.e.*  $\boldsymbol{\theta}$  is a function of  $\mathbf{p} = (p_{00}, \dots, p_{22})^\top$ , the moments estimate  $\hat{\boldsymbol{\theta}}$  is then obtained by plug-in method using  $\hat{\mathbf{p}} = (\hat{p}_{00}, \dots, \hat{p}_{22})^\top$ . With  $n_{x\bullet} = n_{x0} + n_{x1} + n_{x2}$  and  $n_{\bullet y} = n_{0y} + n_{1y} + n_{2y}$ , we get

$$\begin{aligned}\hat{\pi}_1 &= \hat{\theta}_1 = \frac{1}{2N}(n_{1\bullet} + 2n_{2\bullet}), \\ \hat{\pi}_2 &= \hat{\theta}_2 = \frac{1}{2N}(n_{\bullet 1} + 2n_{\bullet 2}), \\ \hat{\rho}_1 &= \hat{\theta}_3 = 1 - \frac{n_{1\bullet}}{2N\hat{\theta}_1(1 - \hat{\theta}_1)}, \\ \hat{\rho}_2 &= \hat{\theta}_4 = 1 - \frac{n_{\bullet 1}}{2N\hat{\theta}_2(1 - \hat{\theta}_2)}, \\ \hat{\rho} &= \hat{\theta}_5 = \frac{1}{4N\hat{\tau}}(n_{11} + 2n_{12} + 2n_{21} + 4n_{22} - 4N\hat{\theta}_1\hat{\theta}_2),\end{aligned}$$

where  $\hat{\tau}^2 = \hat{\theta}_1\hat{\theta}_2(1 - \hat{\theta}_1)(1 - \hat{\theta}_2)$ . Note that  $\hat{\kappa}_1 = (\hat{\theta}_3 - \hat{\theta}_5)/(1 - \hat{\theta}_5)$  and  $\hat{\kappa}_2 = (\hat{\theta}_4 - \hat{\theta}_5)/(1 - \hat{\theta}_5)$ .

From standard asymptotic theory, it follows that  $\hat{\boldsymbol{\theta}}$  has an asymptotic multivariate normal distribution with mean  $\boldsymbol{\theta}$  and covariance matrix  $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = (1/N)(\partial\boldsymbol{\theta}/\partial\mathbf{p})(\mathbf{D} - \mathbf{p}\mathbf{p}^\top)(\partial\boldsymbol{\theta}/\partial\mathbf{p})^\top$ , where  $\mathbf{D} = \text{diag}(\mathbf{p})$  (see online Supporting Information for details). Corresponding estimates  $\hat{P}_{\ell_1 r_1 \ell_2 r_2}$  of the joint probabilities  $P_{\ell_1 r_1 \ell_2 r_2}$  are obtained by plug-in method.

The fit of the model against the general alternative given by the saturated model may be assessed using either the original  $16 \times 16$  table of eye-level counts or the  $3 \times 3$  table of aggregated patient-level counts. In either case, the deviance statistic

$$G^2 = 2 \sum \text{cell count} \times \log \left\{ \frac{\text{cell count}}{N \times \text{cell probability estimate}} \right\},$$

may be used. This statistic has an asymptotic chi-square null distribution with  $N - 5$  degrees of freedom, and  $p$ -value  $\approx P(X_{N-5}^2 > G^2)$ , where  $X_{N-5}^2$  has a chi-square distribution with  $N - 5$  degrees of freedom (Zelterman, 1999, pp. 108–109). As pointed out by a referee, the test based on the aggregated counts may not be sensitive to violation of the assumption  $P_{1010} = P_{0101} = P_{1001} = P_{0110}$  due to possibly good fits for the other probabilities. This may be avoided by carrying out the goodness-of-fit test at the eye-level; however, this may require a fairly large sample size to prevent having cells with small counts.

### 2.2 Measures of diagnostic accuracy

We consider in this section joint estimation of the test’s sensitivity and specificity as well as its positive predictive and negative predictive values. Note that a diagnostic test’s sensitivity and specificity are measures of the test’s intrinsic accuracy and as such, unlike the predictive values, do not provide information on the accuracy of the diagnoses. Because of exchangeability, these measures do not depend on the particular eye under consideration.

The sensitivity  $\text{se} = P(Y_{i1L} = 1 | Y_{i2L} = 1) = P(Y_{i1R} = 1 | Y_{i2R} = 1)$  and specificity  $\text{sp} = P(Y_{i1L} = 0 | Y_{i2L} = 0) = P(Y_{i1R} = 0 | Y_{i2R} = 0)$  of the test are given by

$$\text{se} = \frac{\sum_{r_1=0}^1 \sum_{r_2=0}^1 P_{1r_1 1r_2}}{\sum_{\ell_1=0}^1 \sum_{r_1=0}^1 \sum_{r_2=0}^1 P_{\ell_1 r_1 1r_2}} = \frac{\sum_{\ell_1=0}^1 \sum_{\ell_2=0}^1 P_{\ell_1 1 \ell_2 1}}{\sum_{\ell_1=0}^1 \sum_{r_1=0}^1 \sum_{\ell_2=0}^1 P_{\ell_1 r_1 \ell_2 1}} = \frac{P_{1 \bullet 1 \bullet}}{P_{\bullet \bullet 1 \bullet}} = \frac{P_{\bullet 1 \bullet 1}}{P_{\bullet \bullet 1 \bullet}}, \tag{2}$$

$$\text{sp} = \frac{\sum_{r_1=0}^1 \sum_{r_2=0}^1 P_{0r_1 0r_2}}{\sum_{\ell_1=0}^1 \sum_{r_1=0}^1 \sum_{r_2=1}^1 P_{\ell_1 r_1 0r_2}} = \frac{\sum_{\ell_1=0}^1 \sum_{\ell_2=0}^1 P_{\ell_1 0 \ell_2 0}}{\sum_{\ell_1=0}^1 \sum_{r_1=0}^1 \sum_{\ell_2=0}^1 P_{\ell_1 r_1 \ell_2 0}} = \frac{P_{0 \bullet 0 \bullet}}{P_{\bullet \bullet 0 \bullet}} = \frac{P_{\bullet 0 \bullet 0}}{P_{\bullet \bullet 0 \bullet}}. \tag{3}$$

The positive predictive value  $\text{ppv} = P(Y_{i2L} = 1 | Y_{i1L} = 1) = P(Y_{i2R} = 1 | Y_{i1R} = 1)$  and negative predictive value  $\text{npv} = P(Y_{i2L} = 0 | Y_{i1L} = 0) = P(Y_{i2R} = 0 | Y_{i1R} = 0)$  are given by

$$\text{ppv} = \frac{\text{se}\pi_2}{\text{se}\pi_2 + (1 - \text{sp})(1 - \pi_2)}, \tag{4}$$

$$\text{npv} = \frac{\text{sp}(1 - \pi_2)}{\text{sp}(1 - \pi_2) + (1 - \text{se})\pi_2}. \tag{5}$$

Corresponding estimates are obtained directly by plugging estimates in Section 2 into (2), (3), (4), and (5). The estimate  $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3, \hat{\eta}_4)^\top = (\hat{\text{se}}, \hat{\text{sp}}, \hat{\text{ppv}}, \hat{\text{npv}})^\top$  of  $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \eta_4)^\top = (\text{se}, \text{sp}, \text{ppv}, \text{npv})^\top$  is straightforward to obtain and, unlike those from regression-based methods, it does not require iterative methods to compute. In addition, standard large-sample theory applies for constructing CIs and tests of hypotheses, as  $\hat{\boldsymbol{\eta}}$  has an asymptotic multivariate normal distribution with mean  $\boldsymbol{\eta}$  and covariance matrix  $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}} = (\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta})\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}(\partial\boldsymbol{\eta}/\partial\boldsymbol{\theta})^\top$ . In particular, if  $\eta_k$  is the measure of interest, then a large sample  $(1 - \epsilon)100\%$  CI for  $\eta_k$  is

$$\hat{\eta}_k \pm z_{\epsilon/2} \text{SE}(\hat{\eta}_k), \quad k = 1, \dots, 4, \tag{6}$$

where  $z_{\varepsilon/2}$  the  $(1 - \varepsilon/2)100$ th percentile of the standard normal distribution and  $\text{SE}(\hat{\eta}_k) = \hat{\sigma}_{\hat{\eta}_k}$ , the large-sample standard error of  $\hat{\eta}_k$ , with  $\hat{\sigma}_{\hat{\eta}_k}^2$  the  $k$ th diagonal element of the plug-in estimate  $\hat{\Sigma}_{\hat{\eta}}$  of  $\Sigma_{\hat{\eta}}$ . Note that the calculations involved require no statistical program and can be done in a spreadsheet or by a hand-held calculator. See online Supporting Information for details.

Mercaldo, Lau, and Zhou (2007) note that coverage properties of CIs for proportions, such as the diagnostic accuracy measures considered in this paper, may not be satisfactory due to skewness and non-normality. As a remedy, Mercaldo *et al.* (2007) employed the logit back-transformation method to improve the normal approximation and incorporated the so-called Wilson's continuity correction (Brown, Cai, and DasGupta, 2001) in the estimates of proportions. Adapting their recommendations to the present setting, we can modify the moments estimate  $\hat{\pi}_1$ , for example, by using Wilson's continuity correction, as  $\hat{\pi}_1^w = (n_{1\bullet} + 2n_{2\bullet} + z_{\varepsilon/2}^2/2)/(2N + z_{\varepsilon/2}^2)$ ;  $\hat{\pi}_2$  can be similarly corrected. These modified estimates are then used to obtain the corrected correlation estimates  $\hat{\rho}_1^w, \hat{\rho}_2^w$ , and  $\hat{\rho}^w$  to calculate the Wilson-corrected estimate  $\hat{\eta}^w = (\hat{\eta}_1^w, \hat{\eta}_2^w, \hat{\eta}_3^w, \hat{\eta}_4^w)^\top$  of  $\eta$ . This estimate can be used in lieu of  $\hat{\eta}$ , to construct the standard CI in (6); alternatively, the logit back-transformation method yields the following large-sample  $(1 - \varepsilon)100\%$  CI for  $\eta_k$ :

$$\frac{\exp[\text{logit}(\hat{\eta}_k^w) \pm z_{\varepsilon/2} \text{SE}\{\text{logit}(\hat{\eta}_k^w)\}]}{1 + \exp[\text{logit}(\hat{\eta}_k^w) \pm z_{\varepsilon/2} \text{SE}\{\text{logit}(\hat{\eta}_k^w)\}]}, \quad k = 1, \dots, 4, \quad (7)$$

where  $\text{logit}(\hat{\eta}_k^w) = \log\{\hat{\eta}_k^w / (1 - \hat{\eta}_k^w)\}$  and  $\text{SE}\{\text{logit}(\hat{\eta}_k^w)\} = \hat{\sigma}_{\hat{\eta}_k^w} / \{\hat{\eta}_k^w(1 - \hat{\eta}_k^w)\}$ , with  $\hat{\sigma}_{\hat{\eta}_k^w}^2$  the  $k$ th diagonal element of  $\Sigma_{\hat{\eta}^w}$  evaluated at  $\hat{\eta}^w$ .

Another useful alternative approach to CI construction, which does not rely on large-sample approximations is the bootstrap method (Efron and Tibshirani, 1993). Given  $B$  bootstrap samples from the original data, a percentile method  $(1 - \varepsilon)100\%$  bootstrap CI for  $\eta_k$  is  $[\hat{\eta}_{k,\varepsilon/2}^B, \hat{\eta}_{k,1-\varepsilon/2}^B]$ , where  $\hat{\eta}_{k,\varepsilon/2}^B$  and  $\hat{\eta}_{k,1-\varepsilon/2}^B$  are the respective  $(\varepsilon/2)100$ th and  $(1 - \varepsilon/2)100$ th empirical percentiles of the bootstrap distribution of estimates from model (1). We study the empirical performance of CIs (6) and (7) as well as that of the percentile bootstrap CI in the next section.

### 3 Simulation Study

We investigate in this section the performance of the large-sample and percentile bootstrap Wilson-corrected CIs for  $\eta_k$ ,  $k = 1, \dots, 4$ , in terms of empirical coverage rates and compare them against similar intervals based on GEE. To do this, we carried out a Monte Carlo simulation study. The parameters in the simulations are  $(\pi_1, \pi_2) = (0.5, 0.5), (0.4, 0.4), (0.1, 0.1)$ ,  $\rho_1 = \rho_2 = 0.8, 0.6, 0.5$ ,  $\rho = 0.5, 0.3$ ,  $\varepsilon = 0.05$ , and the number of patients  $N = 100, 250$ . These parameter configurations are relatively common in diagnostic studies in ophthalmology such as the diabetic retinopathy study described earlier. They also define proper joint probability distributions for the binocular data. The number of replications used in the simulations is 2000, which allows for an error margin of about 2.5% for a nominal coverage of 95%, as in Shoukri and Donner (2007). Note that the empirical coverage rates, by jointly incorporating estimates and their associated standard errors as well as the normal approximation, can be used to assess the overall performance of the proposed methodology.

The GEE method (Sternberg and Hadgu, 2001; Smith and Hadgu, 1992) is a non-likelihood-based approach that does not rely on any distributional assumptions regarding the binocular data. In it, reader's diagnoses are marginally regressed on the corresponding disease status as  $P(Y_{i1k} = 1) = g^{-1}(\beta_0 + \beta_1 Y_{i2k})$ ,  $k = L, R$ ,  $i = 1, \dots, N$ , where  $g(t) = \text{logit}(t)$  is the logit link function. Assuming constant intra-pair correlation  $\rho_1$  across patients, the GEE approach relies on a working correlation for  $Y_{i1L}$  and  $Y_{i2R}$  to account for  $\rho_1$ . Standard errors are obtained using the so-called sandwich variance estimate (Smith and Hadgu, 1992). Note that  $\text{se} = g^{-1}(\beta_0 + \beta_1) = e^{\beta_0 + \beta_1} / (1 + e^{\beta_0 + \beta_1})$  and  $\text{sp} = 1 - g^{-1}(\beta_0) = 1 / (1 + e^{\beta_0})$ . Estimates of ppv and npv are obtained from (4) and (5), with  $\pi_2$  estimated by using Wilson's correction. Standard errors are obtained *via* the

delta method, where, for simplicity, it was assumed that the disease-prevalence estimate  $\hat{\pi}_2^w$  is uncorrelated with GEE estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . This implies that the covariance matrix of  $\hat{\pi}_2^w$  and  $(\hat{\beta}_0, \hat{\beta}_1)^T$  is block-diagonal.

For comparison, we considered standard and logit back-transformation CIs studied by Mercaldo *et al.* (2007) and Sternberg and Hadgu (2001). Percentile bootstrap CIs were also constructed based on  $B = 1000$  bootstrap samples using the R package `boot`. Empirical coverage rates were computed as the proportion of simulation repeats a particular CI contained the true parameter values.

Table 1 displays the empirical coverage rates for the 95% model-based Wilson-corrected logit back-transformation CI, the 95% GEE-based Wilson-corrected logit back-transformation CI, and the 95% percentile bootstrap CI; results for the standard CIs are not reported as they show them to be clearly inferior to the logit back-transformation CIs. From Table 1, it appears that logit back-transformation CIs for *se* and *sp* based on moments estimates from model (1) and those based on GEE estimates along with percentile bootstrap CIs have generally similar coverage properties. All three CIs were able to attain the nominal 95% coverage rate (*i.e.* all the empirical levels are within the approximate 95% confidence limits based on the binomial distribution). Table 1 also shows that

**Table 1** Empirical coverage rates of 95% confidence intervals for *se/sp* based on 2000 simulation repeats.<sup>a)</sup>

$\pi_1 = \pi_2$	<i>N</i>	CI	$\rho_1 = \rho_2 = 0.8$		$\rho_1 = \rho_2 = 0.6$		$\rho_1 = \rho_2 = 0.5$	
			$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.3$
0.6	100	A	0.958/0.962	0.959/0.960	0.962/0.956	0.951/0.958	0.938/0.966	0.960/0.962
		B	0.950/0.931	0.953/0.964	0.953/0.936	0.946/0.966	0.946/0.940	0.948/0.966
		C	0.934/0.934	0.939/0.943	0.934/0.941	0.942/0.949	0.946/0.941	0.951/0.952
	250	A	0.952/0.956	0.951/0.956	0.953/0.958	0.949/0.959	0.958/0.960	0.945/0.957
		B	0.952/0.932	0.952/0.960	0.956/0.938	0.952/0.961	0.950/0.941	0.941/0.962
		C	0.945/0.947	0.950/0.951	0.949/0.946	0.947/0.949	0.944/0.946	0.942/0.950
0.5	100	A	0.948/0.951	0.948/0.946	0.949/0.953	0.942/0.952	0.944/0.955	0.947/0.948
		B	0.946/0.968	0.946/0.968	0.956/0.964	0.954/0.965	0.958/0.960	0.952/0.960
		C	0.945/0.946	0.947/0.947	0.943/0.947	0.946/0.947	0.941/0.942	0.940/0.949
	250	A	0.950/0.949	0.944/0.947	0.944/0.957	0.944/0.952	0.942/0.958	0.946/0.951
		B	0.946/0.961	0.946/0.961	0.948/0.966	0.952/0.970	0.950/0.957	0.953/0.961
		C	0.947/0.939	0.953/0.942	0.950/0.943	0.949/0.946	0.946/0.948	0.951/0.952
0.4	100	A	0.960/0.937	0.954/0.940	0.960/0.941	0.958/0.941	0.966/0.937	0.959/0.947
		B	0.948/0.967	0.951/0.950	0.956/0.966	0.951/0.961	0.954/0.968	0.952/0.959
		C	0.936/0.943	0.949/0.949	0.946/0.949	0.944/0.944	0.948/0.956	0.946/0.957
	250	A	0.955/0.941	0.954/0.947	0.958/0.952	0.956/0.947	0.959/0.942	0.955/0.951
		B	0.956/0.961	0.948/0.960	0.951/0.958	0.950/0.960	0.952/0.964	0.950/0.959
		C	0.939/0.956	0.950/0.952	0.951/0.960	0.949/0.953	0.956/0.951	0.951/0.955

a) Confidence intervals (CIs) considered are A: model-based Wilson-corrected logit back-transformation CI, B: GEE-based Wilson-corrected logit back-transformation CI, and C: percentile bootstrap CI. With  $\pi_1 = \pi_2 = 0.6$  and  $\rho_1 = \rho_2 = 0.8, 0.6, 0.5$ , (*se, sp*) = (0.8, 0.7) for  $\rho = 0.5$ , and (*se, sp*) = (0.72, 0.58) for  $\rho = 0.3$ ; with  $\pi_1 = \pi_2 = 0.5$  and  $\rho_1 = \rho_2 = 0.8, 0.6, 0.5$ , (*se, sp*) = (0.75, 0.75) for  $\rho = 0.5$ , and (*se, sp*) = (0.65, 0.65) for  $\rho = 0.3$ ; with  $\pi_1 = \pi_2 = 0.4$  and  $\rho_1 = \rho_2 = 0.8, 0.6, 0.5$ , (*se, sp*) = (0.7, 0.8) for  $\rho = 0.5$ , and (*se, sp*) = (0.58, 0.72) for  $\rho = 0.3$ .

increasing the sample size results in only slight change in the performance of the intervals. These findings confirm earlier results reported by Sternberg and Hadgu (2001).

Results on estimation of ppv and npv via logit back-transformation and percentile bootstrap CIs are displayed in Table 2; again, results for standard CIs are not reported anymore. They indicate that those CIs based on model (1) performed better than those based on GEE, and percentile bootstrap CIs yielded the best performance. While the logit back-transformation CI based on model (1) yielded coverage rates that generally attained, except for a few cases, the desired nominal 95% threshold, the percentile bootstrap CIs achieved the 95% level in all cases considered. Using either moments estimates from model (1) or GEE estimates, the logit back-transformation method resulted in slight improvement in the coverage rates when the sample size is  $N = 250$ . Overall, CIs based on GEE estimates yielded mostly inflated coverage rates, especially for ppv in the case  $\pi_1 = \pi_2 = 0.4$ , *i.e.* low prevalence and low positive diagnosis rate. A possible explanation for this lies in the manner the standard errors were calculated: assuming GEE estimates to be uncorrelated with the prevalence estimate led to overestimation of standard errors resulting in wide CIs. This is

**Table 2** Empirical coverage rates of 95% confidence intervals for ppv/npv based on 2000 simulation repeats.<sup>a)</sup>

$\pi_1 = \pi_2$	$N$	CI	$\rho_1 = \rho_2 = 0.8$		$\rho_1 = \rho_2 = 0.6$		$\rho_1 = \rho_2 = 0.5$	
			$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.3$
0.6	100	A	0.939/0.959	0.943/0.955	0.941/0.972	0.944/0.951	0.951/0.966	0.940/0.958
		B	0.974/0.959	0.954/0.946	0.973/0.960	0.967/0.954	<u>0.978/0.960</u>	<u>0.976/0.964</u>
		C	0.934/0.947	0.931/0.944	0.936/0.947	0.940/0.946	0.931/0.948	0.946/0.950
	250	A	0.946/0.959	0.938/0.955	0.944/0.963	0.942/0.958	0.942/0.968	0.944/0.953
		B	0.972/0.968	0.933/0.946	<u>0.976/0.973</u>	0.947/0.958	<u>0.982/0.978</u>	0.958/0.966
		C	0.948/0.955	0.949/0.955	<u>0.944/0.948</u>	0.954/0.949	0.944/0.948	0.951/0.949
0.5	100	A	0.945/0.939	0.951/0.946	0.946/0.941	0.949/0.943	0.949/0.951	0.947/0.942
		B	<u>0.976/0.973</u>	<u>0.976/0.973</u>	<u>0.984/0.970</u>	0.975/0.970	<u>0.986/0.982</u>	<u>0.978/0.963</u>
		C	0.942/0.955	0.937/0.952	0.940/0.943	0.939/0.943	0.947/0.940	0.944/0.948
	250	A	0.947/0.947	0.949/0.943	0.950/0.944	0.955/0.944	0.956/0.942	0.953/0.949
		B	0.974/0.973	0.974/0.973	<u>0.979/0.969</u>	<u>0.976/0.974</u>	<u>0.976/0.978</u>	<u>0.977/0.966</u>
		C	0.940/0.956	0.948/0.960	<u>0.944/0.954</u>	<u>0.944/0.952</u>	<u>0.956/0.944</u>	<u>0.950/0.951</u>
0.4	100	A	<u>0.977/0.923</u>	0.963/0.930	<u>0.975/0.924</u>	0.966/0.925	<u>0.979/0.976</u>	0.959/0.926
		B	<u>0.986/0.980</u>	<u>0.984/0.962</u>	<u>0.978/0.980</u>	<u>0.988/0.969</u>	<u>0.989/0.978</u>	<u>0.988/0.984</u>
		C	0.938/0.940	0.941/0.951	0.956/0.942	0.942/0.946	0.941/0.956	0.950/0.951
	250	A	0.969/0.930	0.958/0.943	0.974/0.925	0.955/0.936	<u>0.976/0.968</u>	0.958/0.929
		B	<u>0.985/0.981</u>	<u>0.978/0.963</u>	<u>0.979/0.976</u>	<u>0.979/0.964</u>	<u>0.982/0.979</u>	<u>0.979/0.975</u>
		C	<u>0.955/0.944</u>	<u>0.953/0.944</u>	<u>0.949/0.945</u>	<u>0.953/0.947</u>	<u>0.953/0.954</u>	<u>0.950/0.952</u>

a) Confidence intervals (CIs) considered are A: model-based Wilson-corrected logit back-transformation CI, B: GEE-based Wilson-corrected logit back-transformation CI, and C: percentile bootstrap CI. With  $\pi_1 = \pi_2 = 0.6$  and  $\rho_1 = \rho_2 = 0.8, 0.6, 0.5$ , (ppv, npv) = (0.8, 0.7) for  $\rho = 0.5$ , and (ppv, npv) = (0.72, 0.58) for  $\rho = 0.3$ ; with  $\pi_1 = \pi_2 = 0.5$  and  $\rho_1 = \rho_2 = 0.8, 0.6, 0.5$ , (ppv, npv) = (0.75, 0.75) for  $\rho = 0.5$ , and (ppv, npv) = (0.65, 0.65) for  $\rho = 0.3$ ; with  $\pi_1 = \pi_2 = 0.4$  and  $\rho_1 = \rho_2 = 0.8, 0.6, 0.5$ , (ppv, npv) = (0.7, 0.8) for  $\rho = 0.5$ , and (ppv, npv) = (0.58, 0.72) for  $\rho = 0.3$ . Underscored items are those coverage rates that are significantly different from nominal 95% coverage.

clearly remedied by the more computationally intensive percentile bootstrap approach, which yielded excellent coverage rates.

The results in Tables 1 and 2 clearly show the superiority of the percentile bootstrap approach to CI construction, yielding uniformly excellent coverage rates. While admittedly computationally more intensive than the other two methods, the bootstrap method does not require large samples unlike the other two. GEE-based CIs performed comparably well with those based on model (1) for estimating se and sp; however, their performance in estimating ppv and npv was generally inferior, failing to attain the nominal level in many cases. This is because GEE does not allow for simultaneous joint estimation of the four diagnostic accuracy measures. To estimate ppv and npv, GEE requires an extraneous quantity in the form of the prevalence estimate. In case-control studies like those considered by Mercaldo *et al.* (2007), the prevalence rate is assumed known. However, this may not be the case in many cohort or population-based studies. In such situations, a better approach like the percentile bootstrap method, which can account for the correlations between prevalence estimate and GEE estimates, may be used.

### 3.1 Impact of violation of assumption $P_{1010} = P_{0101} = P_{1001} = P_{0110}$

We investigate in this section the impact of violation of the assumption, implicit in model (1), of equal probabilities of perfect agreement and perfect disagreement between diagnosis and disease status for exactly one eye. Specifically, we look into how coverage rates for the CIs are affected whenever  $P_{1010} = P_{0101} \neq P_{1001} = P_{0110}$ .

Let  $a$  be the common value of the probabilities  $P_{1010}$ ,  $P_{0101}$ ,  $P_{1001}$ , and  $P_{0110}$  from model (1). To see how violation of this impacts the CIs, we perturb  $a$  by a small quantity  $\delta > 0$  such that  $P_{1010} = P_{0101} = a - \delta$  and  $P_{1001} = P_{0110} = a + \delta$ . Simulations were carried out with data generated from model (1) with  $\theta = (0.5, 0.5, 0.5, 0.5, 0.5)^T$ . The probabilities  $P_{\ell_1 r_1 \ell_2 r_2}$  were as specified by the model, except that the probabilities of perfect agreement  $P_{1010} = P_{0101}$  and those of perfect disagreement  $P_{1001} = P_{0110}$  were modified as described above. Since  $a = 0.015625$  is small as it is, we chose the following values for the perturbation  $\delta$ : 0.005 (small difference), 0.0075 (moderate difference), and 0.01 (large difference). We used a sample size of  $N = 250$  and 2000 replications in the simulations. Based on the earlier results on coverage rates of various CIs, we decided to compare percentile bootstrap CIs for se, sp, ppv, and npv, based on moments estimates from model (1) and based on GEE estimates. Table 3 displays the coverage rates of the CIs.

Results from Table 3 indicate that violation of the assumption  $P_{1010} = P_{0101} = P_{1001} = P_{0110}$  can greatly affect the coverage rates of CIs based on estimates from model (1). Coverage rates for these CIs were generally lower than the nominal 95% level, and worsened with increasing perturbation  $\delta$ .

**Table 3** Coverage rates for percentile bootstrap CIs for data with  $P_{1010} = P_{0101} = 0.015625 - \delta$  and  $P_{1001} = P_{0110} = 0.015625 + \delta$ .<sup>a)</sup>

	$\delta = 0.005$		$\delta = 0.0075$		$\delta = 0.01$	
	Model	GEE	Model	GEE	Model	GEE
se	0.931	0.944	<u>0.914</u>	0.935	<u>0.886</u>	0.942
sp	0.941	0.942	<u>0.918</u>	0.943	<u>0.893</u>	0.942
ppv	0.939	0.938	<u>0.926</u>	0.934	<u>0.896</u>	0.940
npv	0.936	0.944	<u>0.914</u>	0.940	<u>0.889</u>	0.942

a) Underscored items are those coverage rates that are significantly different from nominal 95% coverage.

This is mainly due to the fact that the moments estimates from model (1) are biased and no longer consistent if  $P_{1010} = P_{0101} \neq P_{1001} = P_{0110}$ , and this bias worsens with  $\delta$ . Percentile bootstrap CIs based on GEE estimates provided excellent coverage rates, however, and attained the 95% threshold in all cases considered. This is not surprising at all since the GEE method, not being likelihood-based, is relatively robust to misspecification of the joint distribution.

The implication of incorrectly assuming equal probabilities  $P_{1010} = P_{0101} = P_{1001} = P_{0110}$  of perfect agreement and perfect disagreement for exactly one eye is clear: failure to account for differences between these probabilities may lead to potentially incorrect inferences. In practice, we suggest that a goodness-of-fit test, like the one described in Section 2, be applied to the  $16 \times 16$  eye-level counts to validate this assumption.

#### 4 Application to Diabetic Retinopathy Data

We now illustrate the proposed methodology on data from a diabetic retinopathy study (de Leon *et al.*, 2007; Rudnisky *et al.*, 2002) involving about a hundred diabetic patients in Alberta, Canada, who were referred to a comprehensive retina practice in Edmonton. The study protocol required that patients be clinically examined on the same day they underwent digital photography by a trained ophthalmic photographer using a high-resolution digital camera. The digital images were stored uncompressed and then graded by an experienced reader at least two months after they were taken. They were assessed in random order, with a minimum of two months in between review of the left eye images and those of the right eyes to minimize reader recall. In order to evaluate treatable diabetic retinopathy among the patients, a number of pathologies that are indicative of retinal thickening were identified as either present (positive) or absent (negative). Contact lens biomicroscopy, the clinical examination considered to be the ‘gold standard,’ was performed on all patients by retinal specialists to determine disease status. Digital images of the patients’ eyes were graded by the reader and patients were diagnosed as either positive or negative for the pathologies.

In what follows, we consider the pathologies macular edema and hard exudate. The former pertains to the thickening and swelling of the eye’s macula due to fluid and protein deposits while the latter involves the leakage of fluid and lipoprotein into the retina of the eye. Table 4 shows data concerning the numbers  $n_{xy}$  of patients with  $y = 0, 1, 2$  eyes with positive status for a pathology and  $x = 0, 1, 2$  eyes diagnosed positive for the pathology by the reader from a total of  $N = 94$  diabetic patients.

Table 5 displays the moments estimates of parameters of model (1) and their large-sample standard errors along with their corresponding percentile bootstrap 95% CIs for macular edema and hard exudate. We note that the inter-eye status correlation estimates of  $\hat{\rho}_2 = 0.685$  and  $\hat{\rho}_2 = 0.623$  for macular edema and hard exudate, respectively, suggest a moderately strong association

**Table 4** Number  $n_{xy}$  of patients with  $y$  status-positive eyes and  $x$  diagnosis-positive eyes.

Macular edema					Hard exudate				
$x$	$y$			$n_{x\bullet}$	$x$	$y$			$n_{x\bullet}$
	0	1	2			0	1	2	
0	42	2	3	47	0	42	4	1	47
1	6	10	2	18	1	5	8	3	16
2	3	2	24	29	2	1	5	25	31
$n_{\bullet y}$	51	14	29	94	$n_{\bullet y}$	48	17	29	94

**Table 5** Model-based estimates, large-sample standard errors (SE), and 95% percentile bootstrap CIs for macular edema and hard exudate.

Parameter	Macular edema			Hard exudate		
	Est.	SE	95% CI	Est.	SE	95% CI
Probability of positive diagnosis	$\pi_1$	0.404	[0.32,0.5]	0.415	0.047	[0.33,0.5]
Prevalence of disease	$\pi_2$	0.383	[0.3,0.48]	0.399	0.046	[0.31,0.48]
<i>Correlations</i>						
Between reader diagnoses	$\rho_1$	0.602	[0.44,0.76]	0.649	0.102	[0.46,0.79]
Between status of fellow eyes	$\rho_2$	0.685	[0.52,0.83]	0.623	0.112	[0.45,0.78]
Aggregate	$\rho$	0.622	[0.48,0.74]	0.681	0.156	[0.45,0.78]
<i>Diagnostic measures</i>						
Sensitivity	se	0.791	[0.68,0.89]	0.81	0.052	[0.72,0.9]
Specificity	sp	0.836	[0.75,0.9]	0.858	0.03	[0.79,0.91]
Positive predictive value	ppv	0.75	[0.63,0.85]	0.792	0.059	[0.69,0.88]
Negative predictive value	npv	0.866	[0.79,0.93]	0.872	0.025	[0.81,0.94]

between the left- and right-eye disease status. Reader diagnoses for left and right eyes are also moderately strongly correlated, as indicated by the correlation estimates  $\hat{\rho}_1 = 0.602$  for macular edema and  $\hat{\rho}_1 = 0.649$  for hard exudate. Note, however, that  $\hat{\rho} = 0.622$  for macular edema and  $\hat{\rho} = 0.681$  for hard exudate, indicating that the associations can be mainly attributed to the aggregated association between left and right eyes.

Estimates of measures of diagnostic accuracy are also shown in Table 5. The sensitivity and specificity estimates for both pathologies range between 79 and 86%. Large-sample standard errors and percentile bootstrap 95% CIs are also shown. These values agree with those previously reported by Rudnisky *et al.* (2002). Goodness-of-fit deviance statistics based on both eye- and patient-level counts yielded non-significant results, indicating that the model provides a reasonably good fit to the data. We also note that there are, respectively, 8/94 and 9/94 observed cases from Table 4 with perfect agreement/disagreement between diagnosis and disease status for exactly one eye, for macular edema and hard exudate, so that the “observed  $a$ ” (in the notation of Section 3.1) is only about 2.5% for both diseases.

A referee correctly pointed out that in practice, it is sufficient that at least one eye is positively diagnosed for the patient to be sent for further and more extensive eye examination. More relevant diagnostic accuracy measures in these cases are the probability of at least one correct positive diagnosis in patients with one or both eyes truly diseased and the probability of two correct negative diagnosis for patients with both eyes truly undiseased. The former is analogous to sensitivity and the latter to specificity. Predictive values may be similarly re-defined. We shall examine these alternative diagnostic measures in a future work.

## 5 Discussion

This paper focused on two main issues arising in reader-based binocular diagnostic studies: how to account for correlation between reader’s diagnoses while at the same time incorporating the correlation between the disease status of fellow eyes. The general approach taken in the paper was a model-based one that relies on specifying a model for the joint distribution of the outcomes. Large-sample inferences based on standard asymptotic theory are then developed for the model. The model is flexible enough to delineate the different associations in the binocular data; in addition, the model yields convenient marginalization properties, and thus can be viewed as a generalization of other simpler commonly used binary data models. The methodology outlined in the paper is illustrated with data from a study concerning retinopathy-related pathologies among diabetic patients. This methodology can be applied not only in ophthalmologic studies but in other contexts as well where paired or binocular data can arise. One such area is twin studies in medical genetics (Shoukri and Donner, 2007).

Methods for constructing CIs for various diagnostic accuracy measures were also investigated and compared in terms of coverage properties. The methods considered incorporated Wilson’s corrections and adopted the logit back-transformation and the percentile bootstrap method to minimize the effects of skewness and nonnormality. The CIs thus obtained were constructed using moments estimates from the proposed model and using GEE estimates. Based on the results of the simulation study, percentile bootstrap CIs gave the most satisfactory performance, with CIs based on the model yielding relatively better coverage properties than those constructed from GEE estimates when the model is correctly specified. This is especially evident in the estimation of ppv and npv, as these require the prevalence rate, a quantity that needs to be estimated extraneously from GEE. In cases where the assumption of equal probabilities of perfect agreement and perfect disagreement  $P_{1010} = P_{0101} = P_{1001} = P_{0110}$  is violated, percentile bootstrap CIs based on GEE estimates provided superior performance over those based on estimates from the model. It is suggested that goodness-of-fit testing be carried out on the eye-level data to check the validity of this assumption.

The approach provides a simple straightforward non-iterative alternative to GEE and other regression-based methods for reader-based binocular diagnostic studies, where diagnostic accuracy measures need to be estimated based on left- and right-eye diagnoses as determined by the same reader. A possible drawback of the model is that the ranges of the correlations may depend on the probabilities, a problem commonly encountered with other binary models (e.g. Bahadur, 1961) as well. Because of this, even if it is possible to theoretically extend the methodology to the multi-reader multi-diseases setting, the joint distribution can become intractable. GEE methods will be a better alternative in these cases.

**Acknowledgements** The authors are grateful to the editor and two anonymous reviewers for their comments. The research of A. R. de Leon is supported by the Natural Sciences and Engineering Research Council of Canada contract/grant number 76-1060 and the University of Calgary Research Grants Committee grant number 78-1045. The authors thank G. Bewa for computational assistance. Computer codes used in the paper and Supporting Information are available on the journal's webpage.

#### Conflict of Interests Statement

The authors have declared no conflict of interest.

## References

- Bahadur, R. R. (1961). A representation of the joint distribution of responses to  $n$  dichotomous items. In: Solomon, H. (Ed.), *Studies in Item Analysis and Prediction*. Stanford University Press, Stanford, pp. 158–168.
- Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16**, 101–133.
- de Leon, A. R., Guo, M., Rudnisky, C. J. and Singh, G. (2007). A likelihood approach to estimating sensitivity and specificity for binocular diagnostic data: application in ophthalmology. *Statistics in Medicine* **26**, 3300–3314.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, Boca Raton.
- Fleiss, J. L., Levin, B. and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions* 3rd edn. Wiley, New York.
- Glynn, R. J. and Rosner, B. (1992). Accounting for the correlation between fellow eyes in regression analysis. *Archives of Ophthalmology* **110**, 381–387.
- Haseman, J. K. and Kupper, L. L. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34**, 69–76.
- Hujoel, P. P., Moulton, L. H. and Loesche, W. J. (1990). Estimation of sensitivity and specificity of sitespecific diagnostic tests. *Journal of Periodontal Research* **25**, 193–196.
- Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions*, Vol. 2. Houghton Mifflin, New York.
- Lefkopoulou, M., Moore, D. and Ryan, L. (1989). The analysis of multiple correlated binary outcomes. *Journal of the American Statistical Association* **84**, 810–815.
- Leisenring, W., Pepe, M. S. and Longton, G. (1997). A marginal regression modelling framework for evaluating medical diagnostic tests. *Statistics in Medicine* **16**, 1263–1281.
- Leite, M. L. C. and Nicolosi, A. (1998). Statistical analysis of correlated binary data in ophthalmology: A weighted logistic regression approach. *Ophthalmic Epidemiology* **5**, 117–131.
- Mak, T. K. (1988). Analysing intraclass correlation for dichotomous variables. *Applied Statistics* **37**, 344–352.
- Mardia, K. V. (1970). *Families of Bivariate Distributions*. Griffin, London.
- Mercaldo, N. D., Lau, K. F. and Zhou, X.-H. (2007). Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine* **26**, 2170–2183.
- Nadarajah, S. and Kotz, S. (2007). Multitude of beta distributions with applications. *Statistics* **41**, 153–179.

- Rosner, B. (1989). Multivariate methods for clustered binary data with more than one level of nesting. *Journal of the American Statistical Association* **84**, 373–380.
- Rudnisky, C. J., Hinz, B. J., Tennant, M. T. S., de Leon, A. R. and Greve, M. D. J. (2002). High-resolution stereoscopic digital fundus photography versus contact-lens biomicroscopy for the detection of clinically significant macular edema. *Ophthalmology* **109**, 267–274.
- Shoukri, M. M. and Donner, A. (2007). Bivariate models for co-aggregation of dichotomous traits in twins. *Statistics in Medicine* **26**, 336–351.
- Shoukri, M. M. and Donner, A. (2009). Bivariate modeling of interobserver agreement coefficients. *Statistics in Medicine* **28**, 430–440.
- Smith, P. J. and Hadgu, A. (1992). Sensitivity and specificity for correlated observations. *Statistics in Medicine* **11**, 1503–1509.
- Sternberg, M. R. and Hadgu, A. (2001). A GEE approach to estimating sensitivity and specificity and coverage properties of the confidence intervals. *Statistics in Medicine* **20**, 1529–1539.
- Sutradhar, B. C. and Das, K. (1997). Generalized linear models for beta correlated binary longitudinal data. *Communications in Statistics–Theory and Methods* **26**, 617–635.
- Zelterman, D. (1999). *Models for Discrete Data*. Oxford University Press, New York.
- Zhou, X.-H., Obuchowski, N. A. and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley, New York.

# Diagnostic Accuracy Measures for Paired Organs—Application in Ophthalmology

*Online Supplementary Materials*

A. R. de LEON<sup>1</sup>, A. SOO<sup>1</sup>, D. BONZO<sup>2</sup>, & C. J. RUDNISKY<sup>3</sup>

<sup>1</sup>*Department of Mathematics & Statistics, University of Calgary*

<sup>2</sup>*Biometrics & Data Management Unit, Xenoport*

<sup>3</sup>*Department of Ophthalmology, University of Alberta*

July 5, 2009

## 1. Asymptotic Normality of $\widehat{\boldsymbol{\theta}}$

Using the fact that the  $n_{xy}$ 's are jointly multinomial with probability vector  $\mathbf{p}$ , by standard asymptotic theory,  $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$  has an asymptotic multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\mathbf{D} - \mathbf{p}\mathbf{p}^\top$ . By the delta method,  $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$  has an asymptotic multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $N\Sigma_{\widehat{\boldsymbol{\theta}}} = (\partial\boldsymbol{\theta}/\partial\mathbf{p})(\mathbf{D} - \mathbf{p}\mathbf{p}^\top)(\partial\boldsymbol{\theta}/\partial\mathbf{p})^\top$ . The elements of  $\partial\boldsymbol{\theta}/\partial\mathbf{p}$  are as follows:

$$\begin{aligned} \frac{\partial\theta_1}{\partial p_{xy}} &= \begin{cases} 0 & , xy = 00, 10, 20 \\ \frac{1}{2} & , xy = 01, 11, 21 \\ 1 & , xy = 02, 12, 22 \end{cases} , \\ \frac{\partial\theta_2}{\partial p_{xy}} &= \begin{cases} 0 & , xy = 00, 01, 02 \\ \frac{1}{2} & , xy = 10, 11, 12 \\ 1 & , xy = 20, 21, 22 \end{cases} , \\ \frac{\partial\theta_3}{\partial p_{xy}} &= \begin{cases} 0 & , xy = 00, 10, 20 \\ -\frac{2-\theta_3}{2\theta_1(1-\theta_1)} & , xy = 01, 11, 21 \\ -\frac{1-\theta_3}{\theta_1(1-\theta_1)} & , xy = 02, 12, 22 \end{cases} , \\ \frac{\partial\theta_4}{\partial p_{xy}} &= \begin{cases} 0 & , xy = 00, 01, 02 \\ -\frac{2-\theta_4}{2\theta_2(1-\theta_2)} & , xy = 10, 11, 12 \\ -\frac{1-\theta_4}{\theta_2(1-\theta_2)} & , xy = 20, 21, 22 \end{cases} , \end{aligned}$$

$$\frac{\partial \theta_5}{\partial p_{xy}} = \begin{cases} 0 & , xy = 00 \\ \frac{1}{2\tau} \left( \theta_5 \sqrt{\frac{\theta_1(1-\theta_1)}{\theta_2(1-\theta_2)}} - \theta_1 \right) & , xy = 10 \\ \frac{1}{2\tau} \left( \theta_5 \sqrt{\frac{\theta_1(1-\theta_1)}{\theta_2(1-\theta_2)}} - 2\theta_1 \right) & , xy = 20 \\ \frac{1}{4\tau} \left( \theta_5 \sqrt{\frac{\theta_2(1-\theta_2)}{\theta_1(1-\theta_1)}} - 2\theta_2 \right) & , xy = 01 \\ \frac{1}{2\tau} \left\{ \theta_5 \left( \sqrt{\frac{\theta_1(1-\theta_1)}{\theta_2(1-\theta_2)}} + \sqrt{\frac{\theta_2(1-\theta_2)}{\theta_1(1-\theta_1)}} \right) \right. \\ \quad \left. + (1 - 2\theta_1 - 2\theta_2) \right\} & , xy = 11 \\ \frac{1}{2\tau} \left\{ \theta_5 \left( \sqrt{\frac{\theta_1(1-\theta_1)}{\theta_2(1-\theta_2)}} + \frac{1}{2} \sqrt{\frac{\theta_2(1-\theta_2)}{\theta_1(1-\theta_1)}} \right) \right. \\ \quad \left. + (1 - 2\theta_1 - \theta_2) \right\} & , xy = 21 \\ \frac{1}{2\tau} \left( \theta_5 \sqrt{\frac{\theta_2(1-\theta_2)}{\theta_1(1-\theta_1)}} - 2\theta_2 \right) & , xy = 02 \\ \frac{1}{2\tau} \left\{ \theta_5 \left( \frac{1}{2} \sqrt{\frac{\theta_1(1-\theta_1)}{\theta_2(1-\theta_2)}} + \sqrt{\frac{\theta_2(1-\theta_2)}{\theta_1(1-\theta_1)}} \right) \right. \\ \quad \left. + (1 - \theta_1 - 2\theta_2) \right\} & , xy = 12 \\ \frac{1}{2\tau} \left\{ \theta_5 \left( \sqrt{\frac{\theta_1(1-\theta_1)}{\theta_2(1-\theta_2)}} + \sqrt{\frac{\theta_2(1-\theta_2)}{\theta_1(1-\theta_1)}} \right) \right. \\ \quad \left. + 2(1 - \theta_1 - \theta_2) \right\} & , xy = 22 \end{cases}.$$

## 2. Partial Derivatives of $P_{\ell_1 r_1 \ell_2 r_2}$

Define the following:

$$\begin{aligned} \phi_j &= E(P_j^{m_j}) = \frac{\Gamma(\alpha_j + m_j)\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\alpha_j + \beta_j + m_j)} = \frac{\Gamma_{j1}\Gamma_{j2}}{\Gamma_{j3}\Gamma_{j4}}, \quad j = 1, 2 \\ \phi_3 &= 1 + \frac{(\gamma_1^{(m_1)} - \pi_1)(\gamma_2^{(m_2)} - \pi_2)}{\rho\tau} = 1 + \frac{\mu_1\mu_2}{\mu_3}, \end{aligned}$$

$\phi_4 = \phi_2\phi_3$ ,  $\phi_5 = \phi_1\phi_3$ , and  $\phi_6 = \phi_1\phi_2$ . Also, let  $f(h) = \frac{\partial f}{\partial \theta_h}$ . Then,

$$\psi_{(h)}(m_1, m_2) = \begin{cases} \phi_{1(h)}\phi_4 + \phi_{2(h)}\phi_5 + \phi_{3(h)}\phi_6 & , m_1 \geq 1, m_2 \geq 1 \\ \phi_{2(h)}\phi_5 + \phi_{3(h)}\phi_6 & , m_1 = 0, m_2 \geq 1 \\ \phi_{1(h)}\phi_4 + \phi_{3(h)}\phi_6 & , m_1 \geq 1, m_2 = 0 \\ \phi_{3(h)}\phi_6 & , m_1 = m_2 = 0 \end{cases},$$

where

$$\begin{aligned} \phi_{j(h)} &= \frac{\Gamma_{j1(h)}\Gamma_{j2} + \Gamma_{j1}\Gamma_{j2(h)}}{\Gamma_{j3}\Gamma_{j4}} - \frac{\Gamma_{j1}\Gamma_{j2}(\Gamma_{j3(h)}\Gamma_{j4} + \Gamma_{j3}\Gamma_{j4(h)})}{(\Gamma_{j3}\Gamma_{j4})^2}, \quad j = 1, 2 \\ \phi_{3(h)} &= \frac{\mu_{1(h)}\mu_2 + \mu_1\mu_{2(h)}}{\mu_3} - \frac{\mu_1\mu_2\mu_{3(h)}}{\mu_3^2}. \end{aligned}$$

We also have for  $j = 1, 2$ ,

$$\begin{aligned}\Gamma_{j1(h)} &= \Gamma(\alpha_j + m_j)\Delta(\alpha_j + m_j)\alpha_{j(h)} \\ \Gamma_{j2(h)} &= \Gamma(\alpha_j + \beta_j)\Delta(\alpha_j + \beta_j)(\alpha_{j(h)} + \beta_{j(h)}) \\ \Gamma_{j3(h)} &= \Gamma(\alpha_j)\Delta(\alpha_j)\alpha_{j(h)} \\ \Gamma_{j4(h)} &= \Gamma(\alpha_j + \beta_j + m_j)\Delta(\alpha_j + \beta_j + m_j)(\alpha_{j(h)} + \beta_{j(h)}),\end{aligned}$$

where  $\Delta(\cdot)$  is the digamma function. We also have the following:

$$\mu_{j(h)} = \begin{cases} -\frac{\rho m_j}{1+\rho(m_j-1)} & , j = h = 1 \text{ or } j = h = 2 \\ 0 & , j = 1, h = 2, 3, 4 \text{ or } j = 2, h = 1, 3, 4 \\ \frac{(m_j - \pi_j)(1 - \pi_j)(1 - \rho)}{\{1 + \rho(m_j - 1)\}^2} & , j = 1, h = 5 \text{ or } j = 2, h = 5 \end{cases}$$

$$\mu_{3(h)} = \begin{cases} \frac{\rho(1-2\pi_1)}{2} \sqrt{\frac{\pi_2(1-\pi_2)}{\pi_1(1-\pi_1)}} & , h = 1 \\ \frac{\rho(1-2\pi_2)}{2} \sqrt{\frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)}} & , h = 2 \\ 0 & , h = 3, 4 \\ \sqrt{\pi_1\pi_2(1-\pi_1)(1-\pi_2)} & , h = 5 \end{cases}.$$

Also, we get

$$\alpha_{j(h)} = \begin{cases} \frac{1-\rho}{\rho} & , j = h = 1 \text{ or } j = h = 2 \\ 0 & , j = 1, h = 2, 3, 4 \text{ or } j = 2, h = 1, 3, 4 \\ -\frac{\pi_j}{\rho^2} & , j = 1, h = 5 \text{ or } j = 2, h = 5 \end{cases}$$

$$\beta_{j(h)} = \begin{cases} -\frac{1-\rho}{\rho} & , j = h = 1 \text{ or } j = h = 2 \\ 0 & , j = 1, h = 2, 3, 4 \text{ or } j = 2, h = 1, 3, 4 \\ -\frac{1-\pi_j}{\rho^2} & , j = 1, h = 5 \text{ or } j = 2, h = 5 \end{cases}.$$

The above expressions can then be used to evaluate partial derivatives for  $P_{\ell_1 r_1 \ell_2 r_2}$ . For example, we get

$$P_{1111(h)} = \begin{cases} \frac{1}{(1-\rho)^2} \left\{ \psi_{(h)}(2, 2)(1 - \rho_1)(1 - \rho_2) + \psi_{(h)}(1, 2)(\rho_1 - \rho)(1 - \rho_2) \right. \\ \quad \left. + \psi_{(h)}(2, 1)(\rho_2 - \rho)(1 - \rho_1) + \psi_{(h)}(1, 1)(\rho_1 - \rho)(\rho_2 - \rho) \right\} & , h = 1, 2 \\ \frac{1}{(1-\rho)^2} \left\{ \psi_{(h)}(2, 2)(1 - \rho_1)(1 - \rho_2) + \psi_{(h)}(1, 2)(\rho_1 - \rho)(1 - \rho_2) \right. \\ \quad - \psi(2, 2)(1 - \rho_2) + \psi(1, 2)(1 - \rho_2) + \psi_{(h)}(2, 1)(\rho_2 - \rho)(1 - \rho_1) \\ \quad \left. - \psi(2, 1)(\rho_2 - \rho) + \psi_{(h)}(1, 1)(\rho_1 - \rho)(\rho_2 - \rho) + \psi(1, 1)(\rho_2 - \rho) \right\} & , h = 3 \\ \frac{1}{(1-\rho)^2} \left\{ \psi_{(h)}(2, 2)(1 - \rho_1)(1 - \rho_2) + \psi_{(h)}(1, 2)(\rho_1 - \rho)(1 - \rho_2) \right. \\ \quad - \psi(2, 2)(1 - \rho_1) - \psi(1, 2)(\rho_1 - \rho) + \psi_{(h)}(2, 1)(\rho_2 - \rho)(1 - \rho_1) \\ \quad \left. + \psi(2, 1)(1 - \rho_1) + \psi_{(h)}(1, 1)(\rho_1 - \rho)(\rho_2 - \rho) + \psi(1, 1)(\rho_1 - \rho) \right\} \\ \quad - \frac{2}{(1-\rho)^3} \left\{ \psi(2, 2)(1 - \rho_1)(1 - \rho_2) + \psi(1, 2)(\rho_1 - \rho)(1 - \rho_2) \right. \\ \quad \left. + \psi(2, 1)(\rho_2 - \rho)(1 - \rho_1) + \psi(1, 1)(\rho_1 - \rho)(\rho_2 - \rho) \right\} & , h = 4 \\ \frac{1}{(1-\rho)^2} \left\{ \psi_{(h)}(2, 2)(1 - \rho_1)(1 - \rho_2) + \psi_{(h)}(1, 2)(\rho_1 - \rho)(1 - \rho_2) \right. \\ \quad - \psi(1, 2)(1 - \rho_2) + \psi_{(h)}(2, 1)(\rho_2 - \rho)(1 - \rho_1) - \psi(2, 1)(1 - \rho_1) \\ \quad \left. + \psi_{(h)}(1, 1)(\rho_1 - \rho)(\rho_2 - \rho) - \psi(1, 1)(\rho_2 - \rho) - \psi(1, 1)(\rho_1 - \rho) \right\} & , h = 5 \end{cases}.$$