STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2384–2399
Published online 4 June 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3627

Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection

Haitao Chu[1,2,3,*,†], Lei Nie[4], Stephen R. Cole[3,5] and Charles Poole[5]

[1]*Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*
[2]*Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*
[3]*Center for AIDS Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*
[4]*Division of Biometrics IV, Office of Biometrics/CDER/OTS/FDA, Spring, MD 20993-0002, U.S.A.*
[5]*Department of Epidemiology, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.*

SUMMARY

In a meta-analysis of diagnostic accuracy studies, the sensitivities and specificities of a diagnostic test may depend on the disease prevalence since the severity and definition of disease may differ from study to study due to the design and the population considered. In this paper, we extend the bivariate nonlinear random effects model on sensitivities and specificities to jointly model the disease prevalence, sensitivities and specificities using trivariate nonlinear random-effects models. Furthermore, as an alternative parameterization, we also propose jointly modeling the test prevalence and the predictive values, which reflect the clinical utility of a diagnostic test. These models allow investigators to study the complex relationship among the disease prevalence, sensitivities and specificities; or among test prevalence and the predictive values, which can reveal hidden information about test performance. We illustrate the proposed two approaches by reanalyzing the data from a meta-analysis of radiological evaluation of lymph node metastases in patients with cervical cancer and a simulation study. The latter illustrates the importance of carefully choosing an appropriate normality assumption for the disease prevalence, sensitivities and specificities, or the test prevalence and the predictive values. In practice, it is recommended to use model selection techniques to identify a best-fitting model for making statistical inference. In summary, the proposed trivariate random effects models are novel and can be very useful in practice for meta-analysis of diagnostic accuracy studies. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; diagnostic tests; sensitivity and specificity; predictive values; sensitivity; specificity

*Correspondence to: Haitao Chu, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, U.S.A.
†E-mail: hchu@bios.unc.edu

Contract/grant sponsor: National Institute of Environmental Health Sciences; contract/grant number: P30ES10126
Contract/grant sponsor: U.S. National Cancer Institute; contract/grant number: CA16086
Contract/grant sponsor: U.S. National Institutes of Health; contract/grant number: P30-AI-50410
Contract/grant sponsor: National Institute of Allergy Infectious Diseases; contract/grant number: R03-AI-071763
Contract/grant sponsor: National Institute of Alcohol Abuse and Alcoholism; contract/grant number: R01-AA-01759

Received 6 June 2008
Accepted 21 April 2009

1. INTRODUCTION

In the presence of a 'gold standard' measure of disease status, the performance of a diagnostic test is often measured by paired indices, such as sensitivity and specificity, positive and negative predictive values, or positive and negative diagnostic likelihood ratios [1, 2]. Sensitivity and specificity are often regarded as intrinsic properties of a diagnostic test. However, it is well understood that sensitivity and specificity may not reflect the clinical utility of a diagnostic test; such clinical utility depends on the prevalence of disease in the population to which the instrument is applied [3]. In particular, high negative predictive value (NPV) or a low negative diagnostic likelihood ratio is necessary for a diagnostic test to be useful at ruling out disease, and high positive predictive value (PPV) or a high positive diagnostic likelihood ratio is necessary for a diagnostic test to be useful at confirming disease. For two diagnostic tests with differing sensitivities and specificities, the PPVs and NPVs may not be strictly ordered because of the nonlinear relationship between PPV and NPV and the triad of sensitivity, specificity and disease prevalence. Specifically, one diagnostic test may have higher PPVs and NPVs for certain levels of disease prevalence, but have lower PPVs and NPVs for other levels of disease prevalence. Such differences may have important implications when comparing the clinical utility of competing diagnostic tests.

There is a great potential for heterogeneity in the meta-analytic studies of diagnostic test accuracy [4–12]. Such heterogeneity arises between studies due to differences in such things as disease prevalence, study population characteristics or laboratory methods. Because of this heterogeneity, random effects models including the hierarchical summary receiver operating characteristic model [4] and bivariate random effects meta-analysis on sensitivities and specificities [6, 10, 12, 13] have been recommended in the literature and by the Cochrane Diagnostic Methods group. Furthermore, Riley *et al.* [14, 15] illustrated that bivariate random-effects meta-analysis offers numerous advantages over separate univariate meta-analysis through extensive simulations. While interest in clinical utility has driven the reporting of positive and negative diagnostic odds ratios, and PPV and NPV in recent meta-analysis of diagnostic test accuracy [16–18], methodological research is sparse on bivariate random effects meta-analysis on positive and negative diagnostic odds ratios, or on PPV and NPV. To our knowledge, only Zwinderman and Bossuyt [11] considered bivariate random effects meta-analysis on positive and negative diagnostic likelihood ratios, and suggested that diagnostic likelihood ratios should not be pooled in systematic reviews. However, meta-analysis on the predictive values using bivariate random effects models has not been described.

Furthermore, although the sensitivity and specificity are often thought of as being independent of disease prevalence for a dichotomous disease status, empirical studies have revealed that this assumption is likely to be violated. One of the reasons is that the classification of disease status is typically based on a continuum of (at least partly) measurable traits. The dependence of sensitivity and specificity with the distribution of the underlying traits, named 'spectrum bias' or 'spectrum effect', was discussed by Ransohoff and Feinstein [19] and others [20, 21]. When the underlying traits for classifying disease status are continuous, subjects with true levels close to the diagnostic test cut-point are more likely to be misclassified. Thus, the underlying distribution of continuous traits not only determines the disease prevalence, but also determines the misclassification rates (i.e. sensitivity and specificity) [22]. Thus, if the underlying distributions of continuous traits are heterogeneous between populations or studies in a meta-analysis, the sensitivities and specificities are more likely to be correlated with the prevalence rates [23].

Given the potential dependence of sensitivities and specificities on the disease prevalence [24], we extend the bivariate random effects model for sensitivity and specificity to jointly model the

DOI: 10.1002/sim

disease prevalence, sensitivity and specificity using a trivariate random effects model, which in return can provide a measurement for the magnitude of the spectrum effect. We further propose to jointly model the diagnostic test prevalence and predictive values as an alternative parameterization to modeling prevalence, sensitivity and specificity. In Section 2, we derive the maximum likelihood function under these two parameterizations, and discuss the estimation of parameters and the selection of random effects. In Section 3, we reanalyze data from a case study of radiological evaluation of lymph node metastases in patients with cervical cancer [25]. We present simulation studies in Section 4 and a discussion in Section 5.

2. STATISTICAL METHODS

We begin with a description of notation. Sensitivity (Se), also referred to as the true positive fraction, is defined as the conditional probability of testing positive in diseased subjects, i.e. $\Pr(T=1|D=1)$ where T and D denote the test and disease status, respectively. Specificity (Sp), also known as the true negative fraction, is defined as the conditional probability of test negative in non-diseased subjects, i.e. $\Pr(T=0|D=0)$. The PPV is defined as the probability of having disease in people with a positive test, i.e. $\Pr(D=1|T=1)$. The NPV is defined as the probability of not having disease in people with a negative test, i.e. $\Pr(D=0|T=0)$.

Typically, the data for each study can be summarized in a 2×2 table as displayed in Table I for study i with n_{i11}, n_{i00}, n_{i01} and n_{i10} denoting the number of true positives, true negatives, false positives and false negatives, $n_{i1}=n_{i11}+n_{i10}$ and $n_{i0}=n_{i01}+n_{i00}$ denoting the number of diseased and healthy subjects, $m_{i1}=n_{i11}+n_{i01}$ and $m_{i0}=n_{i10}+n_{i00}$ denoting the number of test positive and negative subjects and $n_i=n_{i1}+n_{i0}=m_{i1}+m_{i0}$ be the total number of subjects in the ith study, respectively. Let π_i be the study-specific disease prevalence defined as the probability of having disease, i.e. Pr (D=1); P_i be the study-specific test prevalence defined as the probability of being test positive, i.e. Pr (T=1); $\mathrm{Se}_i, \mathrm{Sp}_i, \mathrm{PPV}_i$ and NPV_i be the study-specific sensitivity, specificity, PPV and NPV, respectively. The second and third line in Table I present the likelihood contribution based on the parameterization of $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ and $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, respectively.

2.1. *Random effect model based on parameterization of* π_i, Se_i *and* Sp_i

To take the possible between-study correlation among $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ into consideration, we extend the bivariate generalized linear mixed model approach for diagnostic measures $(\mathrm{Se}_i, \mathrm{Sp}_i)$ [12] to jointly model the disease prevalence, sensitivity and specificity $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$. Specifically, the first stage of the trivariate generalized linear mixed effects model can be specified as follows:

$$n_{i1}|\pi_i \sim \mathrm{Bin}(n_i, \pi_i), \quad n_{i11}|(n_{i1}, \mathrm{Se}_i) \sim \mathrm{Bin}(n_{i1}, \mathrm{Se}_i), \quad n_{i00}|(n_{i0}, \mathrm{Sp}_i) \sim \mathrm{Bin}(n_{i0}, \mathrm{Sp}_i) \tag{1}$$

If $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ are known, the test prevalence, and the PPV and NPV can be estimated from $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ for each study by the following:

$$\begin{aligned} P_i &= \pi_i \mathrm{Se}_i + (1-\pi_i)(1-\mathrm{Sp}_i), \quad \mathrm{PPV}_i = \frac{\pi_i \mathrm{Se}_i}{\pi_i \mathrm{Se}_i + (1-\pi_i)(1-\mathrm{Sp}_i)} \\ \mathrm{NPV}_i &= \frac{(1-\pi_i)\mathrm{Sp}_i}{\pi_i(1-\mathrm{Se}_i) + (1-\pi_i)\mathrm{Sp}_i} \end{aligned} \tag{2}$$

DOI: 10.1002/sim

Table I. Typical data display for the results of a binary diagnostic or screening test compared with a 'gold standard' reference test for study $i\,(i=1,2,\ldots,I)$.

	Gold standard test		
Diagnostic test	Disease (+)	Disease free (−)	Total
Positive (+)	n_{i11} $\pi_i \mathrm{Se}_i$ $\mathrm{PPV}_i \times P_i$	n_{i01} $(1-\pi_i)(1-\mathrm{Sp}_i)$ $(1-\mathrm{PPV}_i)\times P_i$	m_{i1} $\pi_i \mathrm{Se}_i+(1-\pi_i)(1-\mathrm{Sp}_i)$ P_i
Negative (−)	n_{i10} $\pi_i(1-\mathrm{Se}_i)$ $(1-\mathrm{NPV}_i)\times(1-P_i)$	n_{i00} $(1-\pi_i)\mathrm{Sp}_i$ $\mathrm{NPV}_i\times(1-P_i)$	m_{i0} $\pi_i(1-\mathrm{Se}_i)+(1-\pi_i)\mathrm{Sp}_i$ $1-P_i$
Total	n_{i1} π_i $\mathrm{PPV}_i\times P_i+(1-\mathrm{NPV}_i)\times(1-P_i)$	n_{i0} $1-\pi_i$ $(1-\mathrm{PPV}_i)\times P_i+\mathrm{NPV}_i\times(1-P_i)$	n_i 1 1

In each cell, the first line shows the number of subjects, the second and third line present the corresponding probabilities based on the parameterization of $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ and $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, respectively.

To take heterogeneity and potential between-study correlations of the disease prevalence, sensitivities and specificities into consideration, we consider random effects models in the second stage, specified as follows:

$$\mathrm{logit}(\pi_i)=\eta+\varepsilon_i,\quad \mathrm{logit}(\mathrm{Se}_i)=\alpha+\mu_i,\quad \mathrm{logit}(\mathrm{Sp}_i)=\beta+\nu_i \tag{3}$$

where $\mathrm{logit}(p)=\log(p)-\log(1-p)$. The random effects $(\varepsilon_i, \mu_i, \nu_i)^{\mathrm{T}}$ are jointly specified as trivariate normally distributed with mean vector $\mathbf{0}$ and variance–covariance matrix

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_\varepsilon^2 & \rho_{\varepsilon\mu}\sigma_\mu\sigma_\varepsilon & \rho_{\varepsilon\nu}\sigma_\nu\sigma_\varepsilon\\ & \sigma_\mu^2 & \rho_{\mu\nu}\sigma_\mu\sigma_\nu\\ & & \sigma_\nu^2\end{pmatrix}$$

The variance parameters $(\sigma_\varepsilon^2, \sigma_\mu^2, \sigma_\nu^2)$ in the diagonal of the variance–covariance matrix $\boldsymbol{\Sigma}$ capture the between-study heterogeneity of the disease prevalence, sensitivity and specificity, respectively. If there is statistical or scientific evidence of homogeneity among studies (i.e. $\sigma_\varepsilon^2, \sigma_\mu^2, \sigma_\nu^2 \approx 0$), the corresponding study-specific random effects $(\varepsilon_i, \mu_i, \nu_i)$ can be dropped from the model. The parameters $(\rho_{\varepsilon\mu}, \rho_{\varepsilon\nu}, \rho_{\mu\nu})$ capture the corresponding correlation among the random effects. When the correlation parameters $(\rho_{\varepsilon\mu}, \rho_{\varepsilon\nu}, \rho_{\mu\nu})$ are assumed to be zeros, the trivariate model is equivalent to independently fitting three separate univariate models for $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$, respectively. Furthermore, when the correlation parameters $(\rho_{\varepsilon\mu}, \rho_{\varepsilon\nu})$ are assumed to be zeros, the trivariate model is equivalent to independently fitting a univariate model for π_i and a bivariate model for $(\mathrm{Se}_i, \mathrm{Sp}_i)$. To better approximate normality and improve computational performance, the correlation parameters could be transformed by a Fisher's z transformation as $z_i=\frac{1}{2}[\log(1+\rho_i)-\log(1-\rho_i)]$ with $i=1,2,3$ corresponding to the three correlations coefficients $\rho_{\varepsilon\mu}$, $\rho_{\varepsilon\nu}$ and $\rho_{\mu\nu}$.

In the presence of random effects, to obtain the population averages of the disease and test prevalences, sensitivities, specificities and predictive values for all studies in a meta-analysis,

DOI: 10.1002/sim

numerical integration can be implemented over the estimated distributions of random effects. Since the distributions of those parameters are usually highly skewed, the medians are preferred over the means as summary statistics. The medians of disease prevalence, sensitivities and specificities, and an approximation to the medians of test prevalence and predictive values for all studies in a meta-analysis are

$$\pi = [1+\exp(-\eta)]^{-1}, \quad P = [1+\exp(-\eta)]^{-1}[1+\exp(-\alpha)]^{-1} + [1+\exp(\eta)]^{-1}[1+\exp(\beta)]^{-1}$$

$$\mathrm{Se} = [1+\exp(-\alpha)]^{-1}, \quad \mathrm{Sp} = [1+\exp(-\beta)]^{-1}$$

$$\mathrm{PPV} = \frac{\exp(\eta+\alpha)[1+\exp(\beta)]}{\exp(\eta+\alpha)[1+\exp(\beta)]+[1+\exp(\alpha)]} \tag{4}$$

$$\mathrm{NPV} = \frac{\exp(\beta)[1+\exp(\alpha)]}{\exp(\eta)[1+\exp(\beta)]+\exp(\beta)[1+\exp(\alpha)]}$$

Assuming independence among studies conditional on parameters $\boldsymbol{\theta}_i = (\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$, the log likelihood of the observed 2×2 tables conditioning on the random effects $(\varepsilon_i, \mu_i, v_i)$ is the summation of the contribution from each study, that is

$$\begin{aligned}&\sum_i \{n_{i11}\ln(\pi_i \mathrm{Se}_i) + n_{i10}\ln[\pi_i(1-\mathrm{Se}_i)] + n_{i01}\ln[(1-\pi_i)(1-\mathrm{Sp}_i)] + n_{i00}\ln[(1-\pi_i)\mathrm{Sp}_i]\}\\ &= \sum_i \{-n_i \ln[1+\exp(\eta+\varepsilon_i)] - n_{i1}\ln[1+\exp(\alpha+\mu_i)]\\ &\quad -n_{i0}\ln[1+\exp(\beta+v_i)] + n_{i1}(\eta+\varepsilon_i) + n_{i11}(\alpha+\mu_i) + n_{i00}(\beta+v_i)\}\end{aligned} \tag{5}$$

2.2. *Random effect model based on parameterization of* P_i, PPV_i *and* NPV_i

Similar to Section 2.1, we can jointly model the test prevalence, PPVs and NPVs $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$. Specifically, the first stage of the trivariate generalized linear mixed effects model can be specified as follows:

$$m_{i1}|P_i \sim \mathrm{Bin}(n_i, P_i), \quad n_{i11}|(m_{i1}, \mathrm{PPV}_i) \sim \mathrm{Bin}(m_{i1}, \mathrm{PPV}_i)$$

$$n_{i00}|(m_{i0}, \mathrm{NPV}_i) \sim \mathrm{Bin}(m_{i0}, \mathrm{NPV}_i) \tag{6}$$

If $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ are known, the disease prevalence and the sensitivities and specificities can be estimated from $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ for each study by the following:

$$\begin{aligned}\pi_i &= \mathrm{PPV}_i \times P_i + (1-\mathrm{NPV}_i)\times(1-P_i)\\ \mathrm{Se}_i &= \frac{\mathrm{PPV}_i \times P_i}{\mathrm{PPV}_i \times P_i + (1-\mathrm{NPV}_i)\times(1-P_i)}, \quad \mathrm{Sp}_i = \frac{\mathrm{NPV}_i \times (1-P_i)}{(1-\mathrm{PPV}_i)\times P_i + \mathrm{NPV}_i \times (1-P_i)}\end{aligned} \tag{7}$$

To take the heterogeneity and possible correlations of the test prevalence, positive and negative predictive values across studies into account, the random effects model in the second stage can be specified as follows:

$$\mathrm{logit}(P_i) = \eta^* + \varepsilon_\varepsilon^*, \quad \mathrm{logit}(\mathrm{PPV}_i) = \alpha^* + \mu_i^*, \quad \mathrm{logit}(\mathrm{NPV}_i) = \beta^* + v_i^* \tag{8}$$

DOI: 10.1002/sim

with random effects $(\varepsilon_i^*, \mu_i^*, v_i^*)^{\mathrm{T}}$ being specified as trivariate normally distributed with mean vector $\mathbf{0}$ and variance–covariance matrix

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \sigma_\varepsilon^{*2} & \rho_{\varepsilon\mu}^*\sigma_\mu^*\sigma_\varepsilon^* & \rho_{\varepsilon v}^*\sigma_v^*\sigma_\varepsilon^* \\ & \sigma_\mu^{*2} & \rho_{\mu v}^*\sigma_\mu^*\sigma_v^* \\ & & \sigma_v^{*2} \end{pmatrix}$$

The variance parameters $(\sigma_\varepsilon^{*2}, \sigma_\mu^{*2}, \sigma_v^{*2})$ in the diagonal of the variance–covariance matrix $\boldsymbol{\Sigma}$ capture the between-study heterogeneity of the test prevalence, PPVs and NPVs, respectively. The parameters $(\rho_{\varepsilon\mu}^*, \rho_{\varepsilon v}^*, \rho_{\mu v}^*)$ capture the corresponding correlation among the random effects. When the correlation parameters $(\rho_{\varepsilon\mu}^*, \rho_{\varepsilon v}^*, \rho_{\mu v}^*)$ are assumed to be zeros, the trivariate model is equivalent to independently fitting three separate univariate models for P_i, PPV_i and NPV_i, respectively. Similarly, Fisher's z transformations could be applied here to achieve better normality and computational performance for the three correlations coefficients $\rho_{\varepsilon\mu}^*$, $\rho_{\varepsilon v}^*$ and $\rho_{\mu v}^*$.

Similarly, in the presence of random effects and potentially highly-skewed distributions for the parameters. The medians of test prevalence and predictive values, and an approximation to the medians of disease prevalence, sensitivities and specificities for all studies in a meta-analysis are

$$\begin{aligned} \pi &= [1+\exp(-\eta^*)]^{-1}[1+\exp(-\alpha^*)]^{-1}+[1+\exp(\eta^*)]^{-1}[1+\exp(\beta^*)]^{-1} \\ P &= [1+\exp(-\eta^*)]^{-1} \\ \mathrm{Se} &= \frac{\exp(\eta^*+\alpha^*)[1+\exp(\beta^*)]}{\exp(\eta^*+\alpha^*)[1+\exp(\beta^*)]+1+\exp(\alpha^*)} \\ \mathrm{Sp} &= \frac{\exp(\beta^*)[1+\exp(\alpha^*)]}{\exp(\eta^*)[1+\exp(\beta^*)]+\exp(\beta^*)[1+\exp(\alpha^*)]} \\ \mathrm{PPV} &= [1+\exp(-\alpha^*)]^{-1}, \quad \mathrm{NPV} = [1+\exp(-\beta^*)]^{-1} \end{aligned} \tag{9}$$

Assuming independence among studies conditional on parameters $\boldsymbol{\theta}_i^* = (P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, the log likelihood of the observed 2×2 tables conditioning on the random effects $(\varepsilon_i^*, \mu_i^*, v_i^*)$ is the summation of the contribution from each study, that is

$$\begin{aligned} &\sum_i \{n_{i11}\ln(P_i\times\mathrm{PPV}_i)+n_{i10}\ln[(1-P_i)(1-\mathrm{NPV}_i)] \\ &\quad +n_{i01}\ln[P_i(1-\mathrm{PPV}_i)]+n_{i00}\ln[(1-P_i)\mathrm{NPV}_i]\} \\ &= \sum_i \{-n_i\ln[1+\exp(\eta^*+\varepsilon_i^*)]-m_{i1}\ln[1+\exp(\alpha^*+\mu_i^*)] \\ &\quad -m_{i0}\ln[1+\exp(\beta^*+v_i^*)]+m_{i1}(\eta^*+\varepsilon_i^*)+n_{i11}(\alpha^*+\mu_i^*)+n_{i00}(\beta^*+v_i^*)\} \end{aligned} \tag{10}$$

2.3. Estimation of parameters and selection of random effects

To model study-specific covariate effects on the disease prevalence, sensitivity and specificity $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ or alternatively on the test prevalence, PPVs and NPVs $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, we can

DOI: 10.1002/sim

replace the fixed effects (η, α, β) by $(\mathbf{X_i\eta}, \mathbf{Z_i\alpha}, \mathbf{W_i\beta})$ or $(\eta^*, \alpha^*, \beta^*)$ by $(\mathbf{X_i\eta}^*, \mathbf{Z_i\alpha}^*, \mathbf{W_i\beta}^*)$, where $\mathbf{X_i}, \mathbf{Z_i}$ and $\mathbf{W_i}$ are (possibly overlapping) vectors of covariates associated with $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ or $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$.

We adopted a nonlinear mixed effects model to make inference from the above random effects model [26–28], fitted using NLMIXED in SAS version 9.1 (SAS Institute Inc., Cary, NC). NLMIXED maximizes an approximation to the likelihood integrated over the random effects [29]. The adaptive Gaussian quadrature approximation with a maximum of 31 points and dual quasi-Newton algorithm optimization techniques in NLMIXED were used to maximize the approximate integrated likelihood. We computed the population estimates of the disease and test prevalence, sensitivity and specificity, PPVs and NPVs as specified in equations (1) and (2) and their standard errors by the delta method (with the ESTIMATE statement). The study-specific predicted values of disease and test prevalences, sensitivities and specificities, and positive and negative predictive values were computed using empirical-Bayes estimates of the random effects (with the PREDICT statement). In the presence of random effects, estimates of the sensitivity, specificity and PPVs and NPVs represent population medians because distributions of these parameters are usually highly skewed. It is tedious to obtain the population mean estimates, which involve multivariate numerical integration over the estimated distributions of random effects as specified in equations (4) and (9) [30]. Such estimates may be better handled using a Bayesian framework and Monte Carlo methods. To compare goodness-of-fit under various assumptions, we minimize the Akaike's Information Criterion (AIC), which provides a simple way to measure the divergence of the fitted model to the true model [31].

3. EXAMPLE: RADIOLOGICAL EVALUATION OF LYMPH NODE METASTASES IN PATIENTS WITH CERVICAL CANCER

We illustrate the use of the proposed trivariate nonlinear mixed effects model to a meta-analysis of 10 studies of magnetic resonance imaging (MRI) for the diagnosis of lymph node metastasis in women with cervical cancer [25]. The gold standard was surgical pathologic evaluation of tumors, usually performed within several weeks after MRI diagnosis. Our emphasis here is to investigate the impact of different parameterizations, i.e. $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ vs $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, on the estimation of disease and test prevalence, sensitivities and specificities, and PPVs and NPVs of MRI. In this example, the MRI diagnosis was performed before surgical pathologic evaluation and study participants were mainly consecutive patients within a time interval in a particular hospital for each study. Thus, the disease and test prevalence can be considered random for each study allowing us to directly apply the trivariate random effects model.

On the scale of $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$, we consider the trivariate random effects model as in equations (1) and (3), which extends the bivariate meta-analysis of sensitivities and specificities using a nonlinear random effects model [12]. Similarly, on the scale of $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, we consider the trivariate random effects model as in equations (6) and (8).

Table II compares the regression parameter estimates and standard errors obtained from the saturated models (Ia and IIa), the reduced models with zero correlations (Ib and IIb), which are the same as independently fitting three separate univariate random effects models on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ or on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, respectively, and the partially reduced model with $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}) = (0, 0)$ (Ic), which is equivalent to independently fitting a univariate model for π_i and a bivariate model for

DOI: 10.1002/sim

Table II. Summary of parameter estimates with standard errors.

	Model Ia	Model Ib	Model Ic		Model IIa	**Model IIb**
η_0	−1.319 (0.156)	−1.327 (0.162)	−1.327 (0.162)	η_0^*	−1.815 (0.329)	**−1.759 (0.298)**
α_0	0.211 (0.415)	0.157 (0.424)	0.203 (0.423)	α_0^*	0.962 (0.298)	**1.040 (0.266)**
β_0	2.939 (0.332)	2.991 (0.322)	3.046 (0.360)	β_0^*	2.251 (0.235)	**2.152 (0.195)**
σ_ε	0.366 (0.134)	0.391 (0.136)	0.391 (0.136)	σ_ε^*	0.930 (0.300)	**0.838 (0.247)**
σ_μ	1.120 (0.375)	1.150 (0.385)	1.147 (0.388)	σ_μ^*	0.465 (0.309)	**0.416 (0.357)**
σ_v	0.759 (0.303)	0.742 (0.275)	0.871 (0.338)	σ_v^*	0.487 (0.209)	**0.404 (0.187)**
$\rho_{\varepsilon\mu}$	0.588 (0.375)	0	0	$\rho_{\varepsilon\mu}^*$	0.080 (0.715)	**0**
$\rho_{\varepsilon v}$	−0.876 (0.458)	0	0	$\rho_{\varepsilon v}^*$	0.732 (0.332)	**0**
$\rho_{\mu v}$	−0.683 (0.353)	0	−0.735 (0.305)	$\rho_{\mu v}^*$	0.658 (0.670)	**0**
−2logL	1218.6	1223.7	1221.2		1214.6	**1217.9**
AIC	1236.6	1235.7	1235.2		1232.6	**1229.9**

$(\mathrm{Se}_i, \mathrm{Sp}_i)$ for comparison. Based on the AIC, the model with best fit is (IIb) assuming logit normal on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ with zero correlations. Comparing models assuming logit normality on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ (i.e. Ia, Ib and Ic), we note that model (Ic) provides the best-fit, which suggests that, for this example, there is no strong evidence of correlations between disease prevalence and sensitivities or specificities. Although only borderline statistically significant, the correlation coefficients from Model Ia between disease prevalence and sensitivities ($\widehat{\rho}_{\varepsilon\mu} = 0.59$ with a standard error of 0.38 and p-value of 0.12), and disease prevalence specificities ($\widehat{\rho}_{\varepsilon v} = -0.88$ with a standard error of 0.46 and p-value of 0.06) suggest that the sensitivities and specificities of MRI change with the disease prevalence, potentially due to the spectrum effect. Moreover, from Model IIa on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, the correlation coefficient between test prevalence and NPVs ($\widehat{\rho}^*_{\varepsilon v} = 0.73$ with a standard error of 0.33 and p-value of 0.03) suggests that the NPV changes with the test prevalence. On the other hand, the trivial and statistically non-significant correlation coefficients among test prevalence and PPV ($\widehat{\rho}^*_{\varepsilon\mu} = -0.08$ with a standard error of 0.72 and p-value of 0.91) suggest that the PPV is independent of the test prevalence in these data. However, the 10 studies in this meta-analysis is relatively small, which may not provide sufficient statistical power to detect moderate correlations.

Table III presents the population estimates and standard errors of disease and test prevalence, sensitivities and specificities, and PPVs and NPVs from the five models using equations (1) and (3). In general, the two parameterizations give similar but slightly different estimates for the population-averaged parameters. For example, the population median of disease prevalence is estimated to be 0.211 with a standard error of 0.026 if using model Ia, while it is estimated to be 0.183 with a standard error of 0.025 if using model IIa. The population median of negative predictive value is estimated to be 0.886 with standard error of 0.028 using model Ib, while it is estimated to be 0.896 with a standard error of 0.018 using model IIb. These differences are potentially meaningful in practice. Figures 1 and 2 show the 95 per cent confidence region for the summary operating point and a 95 per cent prediction for the operating point in a single future study for sensitivity and specificity (Figure 1(a)), PPV and NPV (Figure 1(b)), disease prevalence and sensitivity (Figure 2(a)), disease prevalence and specificity (Figure 2(b)), test prevalence and PPV (Figure 2(c)), test prevalence and NPV (Figure 2(d)).

DOI: 10.1002/sim

Table III. Summary of disease and test prevalence, sensitivities, specificities and predictive values with standard errors.

	Model Ia	Model Ib	Model Ic	Model IIa	**Model IIb**
Disease prevalence	0.211 (0.026)	0.210 (0.027)	0.210 (0.027)	0.183 (0.025)	**0.197 (0.029)**
Test prevalence	0.156 (0.038)	0.151 (0.028)	0.151 (0.033)	0.140 (0.040)	**0.147 (0.037)**
Sensitivity	0.552 (0.103)	0.539 (0.105)	0.551 (0.105)	0.553 (0.117)	**0.550 (0.087)**
Specificity	0.950 (0.016)	0.952 (0.015)	0.955 (0.016)	0.953 (0.017)	**0.952 (0.016)**
Positive predictive value	0.746 (0.053)	0.749 (0.075)	0.763 (0.061)	0.724 (0.060)	**0.739 (0.051)**
Negative predictive value	0.888 (0.022)	0.886 (0.028)	0.889 (0.027)	0.905 (0.020)	**0.896 (0.018)**

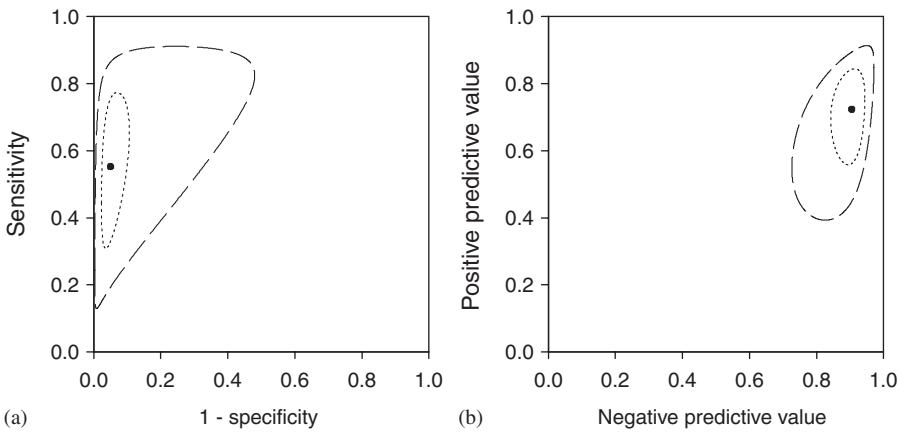

Figure 1. Summary points and regions for sensitivity and specificity (left), positive and negative predictive values (right). Filled circle: summary point; dotted line: boundary of 95 per cent confidence region for the summary point; dashed line: boundary of 95 per cent prediction region.

4. SIMULATION STUDIES

To compare the performance of the trivariate random effects model on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ versus the bivariate random effects model on $(\mathrm{Se}_i, \mathrm{Sp}_i)$, and to study the impact of misspecification of normality assumptions of $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ versus $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ in logit scale, we performed three sets of simulations. In the 1st set of simulations, $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$, in logit scale, is assumed to be trivariate normally distributed with medians of $(0.25, 0.7, 0.9)$, correlation coefficients $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0, 0, 0.7)$, and standard deviations $(\sigma_\varepsilon, \sigma_\mu, \sigma_v) = (0.5, 1.0, 1.0)$ corresponding to a 95 per cent confidence interval of $(0.111, 0.470)$, $(0.247, 0.943)$ and $(0.559, 0.985)$ for $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$. Under these assumptions, the medians of $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ are equal to $(0.25, 0.7, 0.9)$ as well, such that we can easily explore the impact of different assumptions on the estimation. In the 2nd set of simulations, the same parameter values as in the 1st set of simulations are used except the correlation coefficients $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v})$ are $(0.7, 0.7)$. In the 3rd set of simulations, $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, in logit scale, is assumed to be trivariate normally distributed with medians of $(0.25, 0.7, 0.9)$, correlation coefficients $(\rho^*_{\varepsilon\mu}, \rho^*_{\varepsilon v}, \rho^*_{\mu v}) = (0.7, 0.7, 0.7)$, and standard deviations of $(\sigma^*_\varepsilon, \sigma^*_\mu, \sigma^*_v) = (0.5, 1.0, 1.0)$, which corresponds to a 95 per cent confidence interval

DOI: 10.1002/sim

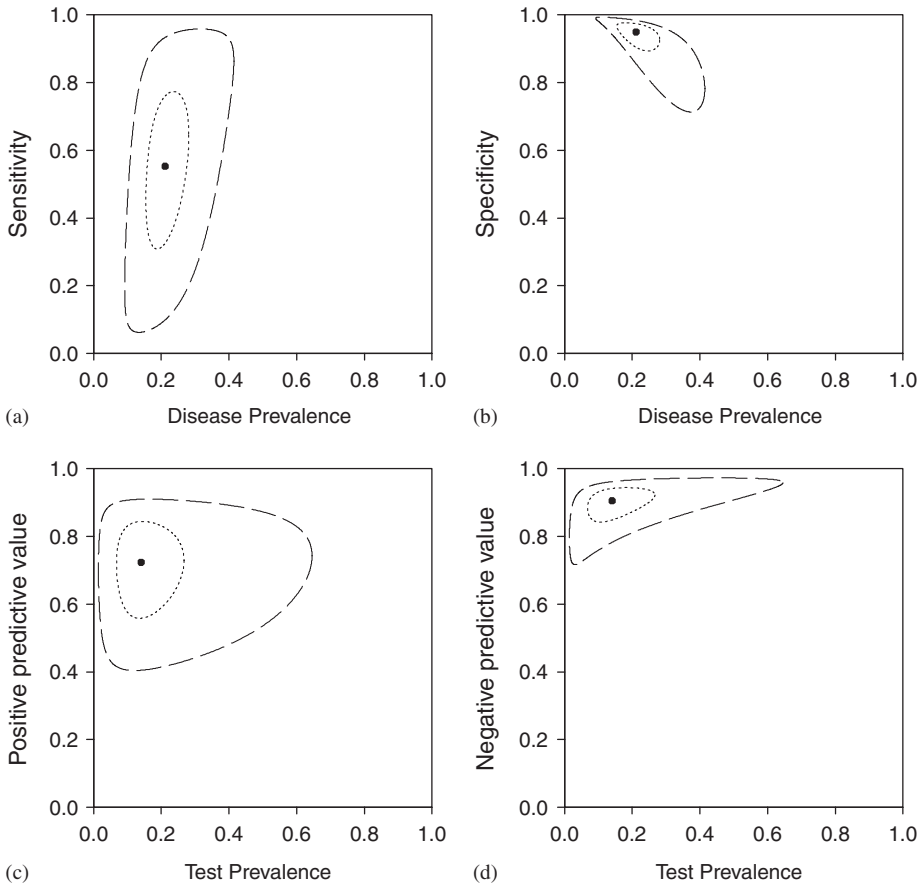

Figure 2. Summary points and regions for disease prevalence and sensitivity (upper left), disease prevalence and specificity (upper right), test prevalence and positive predictive value (bottom left), test prevalence and negative predictive values (bottom right). Filled circle: summary point; dotted line: boundary of 95 per cent confidence region for the summary point; dashed line: boundary of 95 per cent prediction region.

of $(0.111, 0.470)$, $(0.247, 0.943)$ and $(0.559, 0.985)$ for $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$. For each set of simulations, we generated 5000 replicates. For each replicate, 30 studies each with 200 subjects were generated and analyzed by (1) the trivariate model on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ with constrains of correlation parameters $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}) = (0, 0)$, which is equivalent to independently fitting a univariate model for π_i and a bivariate model for $(\mathrm{Se}_i, \mathrm{Sp}_i)$, (2) the trivariate model on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ without any constraints and (3) the trivariate model on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$.

Table IV presents the empirical probabilities of selecting among the three candidate models using AIC based on the 5000 replicates. It shows that the probability of the correct selection is $0.757, 0.906$ and 0.859 when the true models are with logit normal assumption on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ with $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0, 0, 0.7)$ and $(0.7, 0.7, 0.7)$, and logit normal assumption on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ with $(\rho^*_{\varepsilon\mu}, \rho^*_{\varepsilon v}, \rho^*_{\mu v}) = (0.7, 0.7, 0.7)$, respectively. The results suggest that misspecification resulting from AIC-based model selection is reasonably low in this setting. Tables V and VI present

DOI: 10.1002/sim

Table IV. The empirical probability of selecting a candidate link function using AIC based on simulation studies with 5000 replicates. The bolded values represent the probability of identifying the correct model.

	Selected random effects model		
True random effects model	$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ with $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}) = 0$	$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$	$(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$
$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0, 0, 0.7)$	**0.757**	0.141	0.102
$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0.7, 0.7, 0.7)$	0.003	**0.906**	0.091
$(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ $(\rho^*_{\varepsilon\mu}, \rho^*_{\varepsilon v}, \rho^*_{\mu v}) = (0.7, 0.7, 0.7)$	0.100	0.041	**0.859**

the population median estimates, the standard errors and the empirical 95 per cent confidence interval coverage probabilities for $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ and $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, respectively. In summary, the estimated median disease prevalence, sensitivity, specificity, test prevalence, PPV and NPV are unbiased when the correct logit normality is assumed, and slightly biased upon misspecification of logit normality assumption. For example, if data are generated from trivariate logit normal distribution on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ with $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0.7, 0.7, 0.7)$ but fitted with a model assuming trivariate logit normal distribution on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$, the means of population median of disease prevalence, sensitivity and specificity are estimated to be 0.265, 0.738 and 0.880, respectively. Furthermore, the misspecification of the normality assumption and correlation structure has a noticeable impact on the standard error and the 95 per cent confidence interval coverage. Although the confidence interval coverage probabilities are slightly lower than the expected 95 per cent, even if the model is correctly specified, they generally perform well and range from 0.929 to 0.940. However, if the normality assumption is misspecified, a very low coverage probability of 0.404 is observed for the median disease prevalence if we incorrectly specified a trivariate normality assumption on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ when the data are generated based on the trivariate normality assumption on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$; if the correlation structure is misspecified, a low coverage probability of 0.865 is observed for the median PPV if we incorrectly constrained that $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}) = (0, 0)$ even if we correctly specified the normality assumption on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$. It emphasizes the importance to carefully choose an appropriate normality assumption on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ or on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ to make appropriate inferences. Thus, in practice, it may be worth the effort to use AIC or other model selection criteria to select a model that gives the best goodness of fit first, and then to make inference based on the selected model.

5. DISCUSSION

In this paper, we discuss two parameterizations for the estimation of disease and test prevalence, sensitivities, specificities and PPVs and NPVs using trivariate nonlinear random effects models. These models take the heterogeneity across studies into consideration through study-specific random effects on disease prevalence, sensitivities and specificities or on test prevalence, PPVs and NPVs. The two parameterizations make different assumptions on the parameters (i.e. logit-normal on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ versus logit-normal on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$), and due to the nonlinear relationships between $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ and $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ as in equations (2) and (7), the two

DOI: 10.1002/sim

Table V. The population median estimates, the standard errors and the empirical 95 per cent confidence interval coverage probabilities for $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ based on simulation studies with 5000 replicates. The bolded values represent the correctly chosen model.

		True random effects model								
		$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0, 0, 0.7)$			$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0.7, 0.7, 0.7)$			$(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ $(\rho^*_{\varepsilon\mu}, \rho^*_{\varepsilon v}, \rho^*_{\mu v}) = (0.7, 0.7, 0.7)$		
Fitted random effects model		π (0.25)	Se (0.70)	Sp (0.90)	π (0.25)	Se (0.70)	Sp (0.90)	π (0.25)	Se (0.70)	Sp (0.90)
$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ & $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}) = 0$	Mean*	**0.250**	**0.698**	**0.899**	0.250	0.701	0.898	0.285	0.688	0.904
	Standard error	**0.018**	**0.040**	**0.017**	0.018	0.039	0.017	0.016	0.061	0.009
	95 per cent CICP**	**0.939**	**0.929**	**0.928**	0.939	0.929	0.938	0.407	0.932	0.907
$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$	Mean*	0.250	0.698	0.899	**0.250**	**0.698**	**0.899**	0.285	0.688	0.904
	Standard error	0.018	0.040	0.017	**0.018**	**0.040**	**0.017**	0.016	0.061	0.009
	95 per cent CICP**	0.938	0.929	0.929	**0.937**	**0.933**	**0.935**	0.404	0.933	0.906
$(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$	Mean*	0.254	0.720	0.889	0.265	0.738	0.880	**0.251**	**0.695**	**0.900**
	Standard error	0.018	0.037	0.019	0.019	0.034	0.022	**0.014**	**0.059**	**0.010**
	95 per cent CICP**	0.934	0.880	0.932	0.873	0.750	0.890	**0.938**	**0.930**	**0.934**

*Mean of 5000 population median estimates.
**95 per cent CICP=95 per cent confidence interval coverage probability based on normal assumption.

DOI: 10.1002/sim

Table VI. The population median estimates, the standard errors, and the empirical 95 per cent confidence interval coverage probabilities for $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ based on simulation studies with 5000 replicates. The bolded values represent the correctly chosen model.

		True random effects model								
		$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0, 0, 0.7)$			$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}, \rho_{\mu v}) = (0.7, 0.7, 0.7)$			$(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ $(\rho^*_{\varepsilon\mu}, \rho^*_{\varepsilon v}, \rho^*_{\mu v}) = (0.7, 0.7, 0.7)$		
Fitted random effects model		P (0.25)	PPV (0.70)	NPV (0.90)	P (0.25)	PPV (0.70)	NPV (0.90)	P (0.25)	PPV (0.70)	NPV (0.90)
$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ & $(\rho_{\varepsilon\mu}, \rho_{\varepsilon v}) = 0$	Mean*	**0.251**	**0.697**	**0.899**	0.253	0.695	0.901	0.265	0.738	0.880
	Standard error	**0.015**	**0.048**	**0.016**	0.015	0.048	0.016	0.019	0.034	0.023
	95 per cent CICP**	**0.940**	**0.931**	**0.932**	0.959	0.865	0.993	0.851	0.733	0.888
$(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$	Mean*	0.251	0.697	0.899	**0.251**	**0.695**	**0.900**	0.265	0.738	0.880
	Standard error	0.015	0.048	0.016	**0.014**	**0.059**	**0.010**	0.019	0.034	0.022
	95 per cent CICP**	0.938	0.928	0.928	**0.939**	**0.931**	**0.934**	0.873	0.750	0.890
$(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$	Mean*	0.266	0.689	0.903	0.285	0.688	0.904	**0.250**	**0.698**	**0.899**
	Standard error	0.016	0.050	0.015	0.016	0.061	0.009	**0.018**	**0.040**	**0.017**
	95 per cent CICP**	0.845	0.930	0.908	0.404	0.933	0.906	**0.936**	**0.932**	**0.935**

*Mean of 5000 population median estimates.
**95 per cent CICP = 95 per cent confidence interval coverage probability based on normal assumption.

DOI: 10.1002/sim

models are not equivalent. Compared with the bivariate random effects models on sensitivities and specificities, the trivariate random effects models on prevalence and sensitivities and specificities allow investigators to study whether the sensitivities and specificities of a diagnostic test are dependent on the disease prevalence and whether the predictive values of a diagnostic test are dependent on the test prevalence and thus can provide additional information regarding test performance. Depending on the purpose of a meta-analysis, parameterization of disease prevalence, sensitivity and specificity may not be preferred over parameterization of test prevalence, PPV and NPV. Through simulations, we have demonstrated the importance of carefully choosing an appropriate normality assumption on the disease prevalence, sensitivities and specificities, or test prevalence and the predictive values. In practice, it is recommended to use model selection techniques such as AIC to identify a best-fitting model for making statistical inference. In the simulations, we generate data based on a normality assumption either on $(\pi_i, \mathrm{Se}_i, \mathrm{Sp}_i)$ or on $(P_i, \mathrm{PPV}_i, \mathrm{NPV}_i)$ in logit scale, which may not approximate the true diagnostic data generation mechanism. Further research using different distributional assumptions or semi-parametric methods [32, 33] may shed more light on this issue.

In the meta-analysis that we considered, the MRI diagnosis was performed before surgical pathologic evaluation and participants were mainly consecutive patients within a time interval in a particular hospital for each study, i.e. they were not selected based on the status of surgical pathologic evaluation of tumors. This design feature allows us to use the trivariate random effects model to directly investigate the correlations among disease prevalence, sensitivities and specificities, or test prevalence and predictive values. Note that if one or more of the studies in a meta-analysis selected participants based on the disease status using a case-control design, then the disease prevalence based on the observed data for those studies is not representative and thus should not used for the estimation of the distribution of disease prevalence. However, such studies can still be included for the estimation of the distribution of sensitivities and specificities. Our approach can be easily adapted to handle this kind of hybrid meta-analysis with a mixture of cohort and case-control studies by separating the likelihood contributions. That is, those case-control studies only contribute to the bivariate random effects model of sensitivities and specificities, and those cohort studies contribute to the trivariate random effects model of disease prevalence, sensitivities and specificities. Similar strategies can be applied to handle situations where some studies select participants based on the results from diagnostic tests when using the trivariate random effects model for test prevalence and predictive values.

The methods were illustrated using a meta-analysis of 10 studies of MRI evaluation of lymph node metastases in patients with cervical cancer. In this example, the identified moderate correlations among the prevalence, sensitivities and specificities may shed light on the performance of these tests. In the application, we used the delta method, to approximate the standard errors of the population estimates of the disease and test prevalence, sensitivities and specificities and predictive values. An alternative and potentially more accurate method is to estimate the standard errors of those parameters and the corresponding credible intervals using a Bayesian approach, which could be implemented in the free downloadable software WinBUGS [34] and R (http://www.r-project.org/), which may not be directly accessible to most non-statisticians. Moreover, the performance of these two approaches needs to be evaluated for this specific setting by extensive simulation studies. Thus, it is out of the scope for this paper.

It has been shown that the estimation of between-study correlation in the bivariate meta-analysis of diagnostic accuracy studies can be problematic, particularly when the number of studies is relatively small [14], with the correlation often poorly estimated as 1 or -1, causing unstable

DOI: 10.1002/sim

summary parameter estimates. We have implemented a Fisher's Z transformation in the simulation and in the example to normalize the correlation coefficients, which seems to work well in the settings (i.e. 30 studies per meta-analysis) that we considered. Specifically, less than 5 of 5000 replicates did not converge in each set of simulation studies, and the 99th quantile of the estimated correlation coefficients from 5000 replicates ranges from 0.915 to 0.920 for correctly specified models under different scenarios. However, because the trivariate models have three correlation coefficients to estimate, there may be potential estimation problems relating to the correlation parameters, particularly if the sample size is small and the correlation coefficients are large. In practice, since many diagnostic meta-analyses will have less than 30 studies, there may be some computational issues related to the estimation of correlation parameters for applying the trivariate models. Furthermore, if the number of studies in a meta-analysis is small, the AIC may not perform well for suggesting whether we should include correlation parameters in the model since the correlation parameters may be poorly estimated. In this case, one may consider using clinical or substantive knowledge to guide the model selection in addition to the AIC.

It is worthy to note that random effects models have been developed for the estimation of the accuracy of two diagnostic tests in the absence of a 'gold standard' measurement of disease status [35], further statistical methodology research on whether a diagnostic test can be considered as a 'gold standard' measurement is needed, e.g. by comparing the goodness of fit for models assuming no 'gold standard' versus models assuming one of the diagnostic tests is a 'gold standard'. In summary, we have proposed two novel trivariate random effects models for meta-analysis of diagnostic accuracy studies, which can be very useful in practice.

ACKNOWLEDGEMENTS

Dr Poole was supported in part by a grant from the National Institute of Environmental Health Sciences (P30ES10126). Dr Chu was supported in part by the Lineberger Cancer Center Core Grant CA16086 from the U.S. National Cancer Institute and P30-AI-50410 from the U.S. National Institutes of Health. Dr Cole was supported in part by National Institute of Allergy and Infectious Diseases, R03-AI-071763, National Institute of Alcohol Abuse and Alcoholism, R01-AA-01759, and P30-AI-50410. The authors are grateful to the associate editor and referees for their constructive comments and suggestions which have greatly improved this manuscript. This work is completed before Dr Nie joined FDA. Views expressed in this paper are the author's professional opinions and do not necessarily represent the official positions of the U.S. Food and Drug Administration.

REFERENCES

1. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.
2. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
3. Li J, Fine JP, Safdar N. Prevalence-dependent diagnostic accuracy measures. *Statistics in Medicine* 2007; **26**:3258–3273.
4. Rutter CA, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine* 2001; **20**(19):2865–2884.
5. Song FJ, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *International Journal of Epidemiology* 2002; **31**(1):88–95.
6. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589–624.
7. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology* 2004; **57**(9):925–932.

DOI: 10.1002/sim

8. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**(10):982–990.
9. Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *British Medical Journal* 2006; **333**(7565):413.
10. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; **8**(2):239–251.
11. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Statistics in Medicine* 2008; **27**(5):687–697.
12. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006; **59**(12):1331–1332.
13. Chu H, Guo H. Letter to the editor. *Biostatistics* 2009; **10**(1):201–203.
14. Riley R, Abrams K, Sutton A, Lambert P, Thompson J. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**(1):3.
15. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* 2008; **9**(1):172–186.
16. Safdar N, Fine JP, Maki DG. Meta-analysis: methods for diagnosing intravascular device-related bloodstream infection. *Annals of Internal Medicine* 2005; **142**(6):451–466.
17. Pfeiffer CD, Fine JP, Safdar N. Diagnosis of invasive aspergillosis using a galactomannan assay: a meta-analysis. *Clinical Infectious Diseases* 2006; **42**(10):1417–1727.
18. Nishimura K, Sugiyama D, Kogata Y, Tsuji G, Nakazawa T, Kawano S *et al.* Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. *Annals of Internal Medicine* 2007; **146**(11):797–808.
19. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* 1978; **299**(17):926–930.
20. Levy D, Labib SB, Anderson KM, Christiansen JC, Kannel WB, Castelli WP. Determinants of sensitivity and specificity of electrocardiographic criteria for left ventricular hypertrophy. *Circulation* 1990; **81**(3):815–820.
21. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine* 2002; **137**(7):598–602.
22. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine* 1997; **16**(9):981–991.
23. Leeflang MMG, Bossuyt PMM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of Clinical Epidemiology* 2009; **62**(1):5–12.
24. Choi BCK. Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. *Epidemiology* 1997; **8**(1):80–86.
25. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A Meta-analysis. *Journal of the American Medical Association* 1997; **278**(13):1096–1101.
26. Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data* (1st edn). Chapman & Hall, CRC: Boca Raton, 1995.
27. Vonesh EF, Chinchilli VM. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker: New York, 1997.
28. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. Springer: New York, 2005.
29. Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**(1):12–35.
30. Halloran ME, Preziosi MP, Chu HT. Estimating vaccine efficacy from secondary attack rates. *Journal of the American Statistical Association* 2003; **98**(461):38–46.
31. Kullback S, Leibler RA. On information and sufficiency. *Annals of Mathematical Statistics* 1951; **22**:79–86.
32. Liang KY, Zeger SL. Longitudinal data-analysis using generalized linear-models. *Biometrika* 1986; **73**(1):13–22.
33. Zeger SL, Liang KY. Longitudinal data-analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**(1): 121–130.
34. Spiegelhalter DJ, Thomas A, Best NG. *WinBUGS User Manual*, *Version 1.4*, 2002. Ref Type: unpublished work.
35. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. *Journal of the American Statistical Association* 2009; DOI: 10.1198/jasa.2009.0017