A hybrid model for combining case–control and cohort studies in systematic reviews of diagnostic tests

Yong Chen and Yulun Liu,
University of Texas School of Public Health, Houston, USA

Jing Ning and Janice Cormier
University of Texas MD Anderson Cancer Center, Houston, USA

and Haitao Chu
University of Minnesota School of Public Health, Minneapolis, USA

[Received February 2013. Final revision July 2014]

Summary. Systematic reviews of diagnostic tests often involve a mixture of case–control and cohort studies. The standard methods for evaluating diagnostic accuracy focus only on sensitivity and specificity and ignore the information on disease prevalence that is contained in cohort studies. Consequently, such methods cannot provide estimates of measures related to disease prevalence, such as population-averaged or overall positive and negative predictive values, which reflect the clinical utility of a diagnostic test. We propose a hybrid approach that jointly models the disease prevalence along with diagnostic test sensitivity and specificity in cohort studies, and sensitivity and specificity in case–control studies. To overcome the potential computational difficulties in the standard full likelihood inference of the hybrid model proposed, we propose an alternative inference procedure based on composite likelihood. Such composite-likelihood-based inference does not suffer computational problems and maintains high relative efficiency. In addition, it is more robust to model misspecifications compared with standard full likelihood inference. We apply our approach to a review of the performance of contemporary diagnostic imaging modalities for detecting metastases in patients with melanoma.

Keywords: Composite likelihood; Diagnostic accuracy study; Independence likelihood; Multivariate meta-analysis; Prevalence

1. Introduction

Comparative effectiveness research relies fundamentally on the accurate assessment of clinical outcomes. Rapid escalations in the cost of medical diagnostic tests, together with growth in the number of available instruments, have generated an increasing need for scientifically rigorous methods for comparing diagnostic tests in clinical practice. Many quantitative comparisons of diagnostic tests are based on systematic reviews of diagnostic accuracy. In such reviews, the performance of a diagnostic test is often summarized by paired indices, such as sensitivity and specificity, or positive and negative predictive values (PPV and NPV) (Pepe, 2004; Zhou et al., 2009).

Address for correspondence: Yong Chen, Division of Biostatistics, University of Texas School of Public Health, 1200 Pressler Street, Houston, TX 77030, USA.
E-mail: yong.chen@uth.tmc.edu

© 2014 Royal Statistical Society
The procedure of pooling paired indices is not straightforward because of three important characteristics of such data. The first is that the estimated sensitivities and specificities are typically negatively correlated between studies (Reitsma et al., 2005). A possible cause of this negative correlation is that studies may have used different thresholds to define positive and negative test results. The second important characteristic is the substantial between-study heterogeneity in paired indices (Moses et al., 1993; Irwig et al., 1995; Rutter and Gatsonis, 2001). Such heterogeneity may arise from differences in study population characteristics, variability of assessment and other factors. The third characteristic is that some prevalence-dependent indices, such as predictive values, i.e. PPV and NPV, require knowledge about disease prevalence that is not estimable in case–control studies. In addition, it has been suggested that sensitivity and specificity can be correlated with disease prevalence. One of the reasons is that the classification of disease status is typically based on a continuum of measurable traits. For classification of disease status based on continuous traits, the underlying distribution of the continuous traits not only determines disease prevalence but also determines misclassification rates (i.e. sensitivity and specificity), because subjects with true levels close to the cut point are more likely to be misclassified (Leefflang et al., 2008; Chu et al., 2009). If the underlying distributions of continuous traits are heterogeneous across studies, the sensitivities and specificities are likely to be correlated with the prevalence.

A variety of methods have been proposed to account for the first two characteristics of data encountered in systematic reviews of diagnostic test accuracy; see the recent review by Ma et al. (2014). The current methods can be classified into two categories. The first category consists of methods based on a summary receiver operating characteristic curve generated from the study data (Littenberg and Moses, 1993; Moses et al., 1993) and a hierarchical summary receiver operating characteristic model (Rutter and Gatsonis, 1995, 2001; Walter, 2002; Arends et al., 2008). The second category consists of methods that use bivariate general mixed effects models and bivariate generalized linear mixed models (GLMMs) to describe sensitivity and specificity simultaneously (Van Houwelingen et al., 1993, 2002; Reitsma et al., 2005; Chu and Cole, 2006; Arends et al., 2008; Hamza et al., 2008; Chu et al., 2010). Interestingly, the hierarchical summary receiver operating characteristic model and the bivariate GLMM have been found to be very closely related, and even identical in the absence of covariates (Harbord et al., 2007; Chu and Guo, 2009).

More recently, a trivariate GLMM has been proposed that can simultaneously account for all three aforementioned characteristics of the data in a systematic review of diagnostic test accuracy (Chu et al., 2009). The trivariate GLMM jointly models disease prevalence with diagnostic test sensitivity and specificity based on the data from the cohort studies. On the basis of the estimated disease prevalence, the clinically meaningful indices, such as PPV and NPV, are immediately available. However, the trivariate GLMM also has a few limitations. First, systematic reviews often involve a mixture of case–control and cohort studies, whereas a trivariate GLMM is restricted to cohort studies. In this situation, discarding all of the case–control studies can lead to a substantial loss of efficiency. Secondly, the correlations between disease prevalence, sensitivity and specificity need to be estimated, and three-dimensional integrals must be evaluated in the likelihood. In this case, non-convergence problems and singular information matrices have been reported when maximizing the likelihood, especially when the number of studies is relatively small (Hamza et al., 2008). Although modern computational techniques such as Laplace or adaptive Gaussian quadrature approximation are available in software, such as procedure NL MIXED in SAS (SAS Institute, Cary) and the automatic differentiation model builder, these approximations may still have non-negligible approximation errors and the estimates may be sensitive to initial values, leading to unstable or unrepeatable
results. To the best of our knowledge, there is currently no satisfactory solution to these limitations.

Our motivating study is a systematic review of modern diagnostic imaging modalities for surveillance of melanoma patients. Melanoma is the least common but most deadly type of skin cancer and occurs in melanocytes, which are cells that produce the skin pigment melanin. Melanoma accounts for more than 75% of deaths related to skin cancer (Jerant et al., 2000). Sentinel lymph node biopsy is the gold standard for pathological staging of metastasis in melanoma. Diagnostic imaging is often utilized following the surgical treatment of melanoma in patients who are at high risk of recurrence of disease. The type of imaging and the interval of testing which are the most effective and cost effective have not been defined. The goal of surveillance imaging is to detect recurrence of melanoma in regional lymph nodes and/or distant sites at a point when it remains treatable and/or possibly surgically resectable. Current diagnostic imaging modalities for the surveillance of melanoma patients include ultrasonography, computed tomography (CT), positron emission tomography (PET) and a combination of both (PET–CT). It is critical to assess and compare the performance of these contemporary diagnostic imaging modalities to compare accuracy in various clinical settings and to support clinical decision making. A systematic review of published studies has examined diagnostic modality characteristics and identified 98 studies from 10 528 patients with melanoma between January 1st, 1990, and June 30th, 2009 (Xing et al., 2011). Out of 98 studies, 57 were cohort studies and the numbers of case–control and cohort studies stratified by type of cancer and type of imaging modality are summarized in Table S1 in the on-line supplementary materials. The original analysis in Xing et al. (2011) treated the case–control studies and cohorts equivalently and ignored the information on prevalence of melanoma. Consequently, clinically more relevant measures, such as the overall PPV and NPV, cannot be obtained.

In this paper, we propose a hybrid multivariate random-effects model to combine case–control and cohort studies. Such a strategy of joint modelling fully utilizes the data and can provide estimates of measures related to disease prevalence (e.g. PPV and NPV). However, the standard likelihood-based inference of the hybrid model proposed is still subject to the aforementioned non-convergence problem and computational difficulty. Motivated by the fact that the commonly used measures of diagnostic tests (e.g. sensitivity, specificity, PPV and NPV) do not involve correlation parameters, we propose an alternative inference procedure based on composite likelihood (CL) (Lindsay, 1988; Chandler and Bate, 2007) where a working independence assumption is adopted. Simulation studies suggest that CL, which avoids an explicit modelling of the dependence structure, does not lead to a substantial loss of efficiency. Therefore, CL inference is a practical solution to the non-convergence problem and computational difficulty. Furthermore, the inference based on CL relies on only the marginal normality of logit prevalence, sensitivity and specificity. Hence the CL method can be more robust than standard full likelihood (FL) inference to misspecifications of the joint distribution. In fact, CL has been widely applied to applications such as longitudinal data analysis and multivariate survival data analysis (Henderson and Shimakura, 2003; Fieuws and Verbeke, 2006; Barry and Bowman, 2008; Molenberghs and Verbeke, 2005). However, to the best of our knowledge, the present paper is the first application of CL in systematic reviews of diagnostic tests.

This paper is organized as follows. In Section 2, we describe the hybrid model proposed and two inference procedures, namely the FL and CL methods. In Section 3, we conduct simulation studies to compare these two inference procedures. We apply our method in Section 4 to a systematic review of the accuracy of contemporary diagnostic imaging modalities for detecting metastases in patients with melanoma. We provide a brief discussion in Section 5.
The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

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

2. Statistical methodology

Denote $D$ and $T$ as the respective disease status ascertained by a gold standard and the result from a diagnostic test under investigation (1, positive; 0, negative). Sensitivity $Se$ and specificity $Sp$ are respectively the probability of a positive test result in a subject with the disease and the probability of a negative test result in a subject who does not have the disease, i.e. $Se = \Pr(T = 1|D = 1)$ and $Sp = \Pr(T = 0|D = 0)$. The positive predictive value PPV and negative predictive value NPV are the probability of having the disease given a positive test result and the probability of not having the disease given a negative test result, i.e. $PPV = \Pr(D = 1|T = 1)$ and $NPV = \Pr(D = 0|T = 0)$ respectively.

We consider a systematic review of diagnostic test accuracy with $m$ studies. For simplicity of notation, assume that the first $m_1$ studies are case–control studies and the remaining $m_2$ studies are cohort studies ($m = m_1 + m_2$). Table 1 summarizes typical data from the $i$th study by a $2 \times 2$ table (Honest and Khan, 2002). Specifically, denote $n_{i11}, n_{i00}, n_{i01}$ and $n_{i10}$ as the number of true positive, true negative, false negative and false positive results respectively, and denote $n_{i1} = n_{i11} + n_{i01}$ and $n_{i0} = n_{i10} + n_{i00}$ as the numbers of subjects with and without the disease respectively. Let $\pi_i$, $Se_i$ and $Sp_i$ be the study-specific disease prevalence, and diagnostic test sensitivity and specificity respectively. Note that $\pi_i$ is estimable only in cohort studies, i.e. $i = m_1 + 1, \ldots, m$.

In practice, there is often significant heterogeneity in the study-specific disease prevalence, and test sensitivity and specificity across studies due to differences in study population characteristics, assessment methods and intervals, and other related factors. Additionally, in practice, diagnostic test sensitivity and specificity are often negatively correlated, and such sensitivity and specificity can be correlated with disease prevalence in cohort studies (Leeflang et al., 2008; Chu et al., 2009). To account for these characteristics and to combine case–control and cohort stud-

| Table 1. Possible data outcomes and probabilities for study $i$ ($i = 1, \ldots, m$)† |
|-------------------------------|-------------------|-------------------|
| **Diagnostic test (T)**       | **Disease status by a gold standard test (D)** |          |
|                               | **Case–control studies** (i = 1, 2, \ldots, $m_1$) | **Cohort studies** (i = $m_1 + 1, m_1 + 2, \ldots, m$) |
|                               | Disease (+)       | Non-disease (−)   | Disease (+)       | Non-disease (−)   |
| Positive (T+)                 | $n_{i11}$         | $n_{i10}$         | $n_{i11}$         | $n_{i10}$         |
| $Se_i$                        | 1 − $Sp_i$        | $\pi_iSe_i$       | (1 − $\pi_i$)(1 − $Sp_i$) |
| Negative (T−)                 | $n_{i01}$         | $n_{i00}$         | $n_{i01}$         | $n_{i00}$         |
| 1 − $Se_i$                    | $Sp_i$            | $\pi_i(1 − Se_i)$ | (1 − $\pi_i$)$Sp_i$ |
| Total                         | $n_{i1}$          | $n_{i0}$          | $n_{i1}$          | $n_{i0}$          |
| 1                             | 1                 | $\pi_i$           | 1 − $\pi_i$       |

†In each cell, the first row shows the observed count and the second row shows the corresponding conditional probability of test outcome given disease status in case–control studies, or the corresponding probability of cell memberships in cohort studies.
ies effectively, we propose the following hybrid generalized linear mixed effects model (which is referred to as the hybrid model hereafter). This model can be formulated in two stages. The first stage specifies the probability of observing the data that are described in Table 1 for a given study-specific disease prevalence (for a cohort study only), and diagnostic test sensitivity and specificity: for \(i = 1, \ldots, m_1\) (i.e. case–control studies),

\[
(n_{i11}, n_{i10}, n_{i01}, n_{i00})|\left(\pi_i, Se_i, Sp_i\right) \\
\sim \text{binomial}(n_{i11}; Se_i) \times \text{binomial}(n_{i00}; Sp_i);
\]

for \(i = m_1 + 1, \ldots, m\) (i.e. cohort studies),

\[
(n_{i11}, n_{i10}, n_{i01}, n_{i00})|\left(\pi_i, Se_i, Sp_i\right) \\
\sim \text{multinomial}\{n_i; \pi_i, Se_i, (1 - \pi_i)(1 - Sp_i), \pi_i(1 - Se_i), (1 - \pi_i) Sp_i\},
\]

where \(n_i\) is the number of subjects in the \(i\)th study, and binomial(\(-\cdot\)) and multinomial(\(-\cdot\)) are defined as

\[
\text{binomial}(y; n, p) = \binom{n}{y} p^y (1 - p)^{n-y}
\]

and

\[
\text{multinomial}(y_1, y_2, y_3, y_4; n, p_1, p_2, p_3, p_4) = \binom{n}{y_1, y_2, y_3, y_4} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4},
\]

where \(p_1 + p_2 + p_3 + p_4 = 1\) and \(y_1 + y_2 + y_3 + y_4 = n\). At the second stage, a random-effects model is assumed to take into consideration the heterogeneity between studies, the correlation between \((Se_i, Sp_i)\) for \(i = 1, \ldots, m_1\) and the correlations between \((\pi_i, Se_i, Sp_i)\) for \(i = m_1 + 1, \ldots, m\): for \(i = 1, \ldots, m_1\) (i.e. case–control studies),

\[
g(Se_i) = X_i^T \beta_1 + \mu_{i1},
\]

\[
g(Sp_i) = Z_i^T \beta_2 + \mu_{i2};
\]

for \(i = m_1 + 1, \ldots, m\) (i.e. cohort studies),

\[
g(\pi_i) = W_i^T \beta_0 + \mu_{i0},
\]

\[
g(Se_i) = X_i^T \beta_1 + \mu_{i1},
\]

\[
g(Sp_i) = Z_i^T \beta_2 + \mu_{i2},
\]

where \(g(\cdot)\) is a known link function such as a logit function and \(W_i, X_i\) and \(Z_i\) are vectors of study level covariates, possibly overlapping, related to \(\pi_i, Se_i\) and \(Sp_i\) respectively. Examples of such study level covariates include the type of disease (e.g. regional cancer \emph{versus} distant cancer) and the quality assessment of diagnostic accuracy studies scale (Whiting et al., 2003). Here we model specificity (i.e. Pr(\(T = 0|D = 0\)) instead of the probability Pr(\(T = 1|D = 0\)) because correctly identifying the disease status (i.e. \(T = 0\)) given a patient without disease (i.e. \(D = 0\)) is considered as a 'success event'. The random intercepts \((\mu_{i1}, \mu_{i2})\) for a case–control study and \((\mu_{i0}, \mu_{i1}, \mu_{i2})\) for a cohort study are assumed respectively to follow a bivariate normal distribution with mean 0 and covariance matrix \(\Sigma_1\) for \(i = 1, \ldots, m_1\) and a trivariate normal distribution with mean 0 and covariance matrix \(\Sigma_2\) for \(i = m_1 + 1, \ldots, m\), defined as

\[
\Sigma_1 = \begin{pmatrix}
\tau_1^2 & \rho_{12}\tau_1\tau_2 \\
\rho_{12}\tau_1\tau_2 & \tau_2^2
\end{pmatrix}
\]

and
The parameters \( \tau_0^2, \tau_1^2 \) and \( \tau_2^2 \) capture the between-study heterogeneity in disease prevalence and test sensitivities and specificities respectively, and the parameters \( \rho_{01}, \rho_{02} \) and \( \rho_{12} \) describe the correlations between the respective random effects \((\pi_i, \text{Se}_i), (\pi_i, \text{Sp}_i)\) and \((\text{Se}_i, \text{Sp}_i)\) in the transformed scale respectively.

To simplify the notation and to make our discussion concrete, we assume that \( W_i = X_i = Z_i = 1 \) and choose a logit link function. In this case, \( \beta_0, \beta_1 \) and \( \beta_2 \) specify the overall disease prevalence, and diagnostic test sensitivity and specificity (on the logit scale) respectively. Besides sensitivity and specificity, other clinically relevant measures, e.g. the overall PPV and NPV, can be calculated as

\[
\begin{align*}
\text{PPV} &= \frac{\exp(\beta_0 + \beta_1)\{1 + \exp(\beta_2)\}}{\exp(\beta_0 + \beta_1)\{1 + \exp(\beta_2)\} + 1 + \exp(\beta_1)}, \\
\text{NPV} &= \frac{\exp(\beta_2\{1 + \exp(\beta_1)\}) + \exp(\beta_0\{1 + \exp(\beta_2)\})}{\exp(\beta_2\{1 + \exp(\beta_1)\}) + \exp(\beta_0\{1 + \exp(\beta_2)\})}.
\end{align*}
\]

In practice, a high NPV is required for a diagnostic test to be useful in ruling out disease, and a high PPV is required for a diagnostic test to be useful in detecting disease.

For simplicity of notation, denote \( \theta_0 = (\beta_0, \tau_0^2)^T, \theta_1 = (\beta_1, \tau_1^2)^T, \theta_2 = (\beta_2, \tau_2^2)^T \) and \( \rho = (\rho_{01}, \rho_{02}, \rho_{12})^T \). Under the hybrid model assumption, the log-likelihood function is

\[
\log L(\theta_0, \theta_1, \theta_2, \rho) = \sum_{i=1}^{m_1} \log \left\{ \int \int \text{binomial}(n_{i00}, n_{i11}; \phi_1) \times \text{binomial}(n_{i11}, n_{i10}; \phi_2) \times \phi_1 (\text{Se}_i, \text{Sp}_i; \theta_1, \theta_2, \rho_1) \text{dSe}_i \text{dSp}_i \right\} \\
+ \sum_{i=m_1+1}^{m} \log \left[ \int \int \int \text{multinomial}(n_{i11}, n_{i10}, n_{i01}; \pi_i, \text{Se}_i, (1 - \pi_i)(1 - \text{Sp}_i), \pi_i(1 - \text{Se}_i)) \times \phi_2 (\pi_i, \text{Se}_i, \text{Sp}_i; \theta_0, \theta_1, \theta_2, \rho_i) \text{dSe}_i \text{dSp}_i \right],
\]

where \( \phi_1(\cdot, \cdot; \theta_1, \theta_2, \rho_1) \) is the bivariate logit normal density with mean \((\beta_1, \beta_2)^T\) and covariance matrix \( \Sigma_1 \) and \( \phi_2(\cdot, \cdot, \cdot; \theta_0, \theta_1, \theta_2, \rho) \) is the trivariate logit normal density with mean \((\beta_0, \beta_1, \beta_2)^T\) and covariance matrix \( \Sigma_2 \) on the logit scale. The integrals in equation (4) do not have a closed form and must to be calculated by using numerical methods such as adaptive Gaussian quadrature (Pinheiro and Bates, 1995). In practice, the package \texttt{NLMIXED} in SAS version 9.3 can be used to maximize the approximation to the log-likelihood function in equation (4). Other methods such as the automatic differentiation model builder (http://admb-project.org) can also be used to approximate this likelihood.

Although conceptually straightforward, the standard maximum FL method faces the non-convergence and computational problems as described in Section 1. These problems are due to the two- and three-dimensional integrals in the likelihood function \( L(\theta_0, \theta_1, \theta_2, \rho) \) and the need for estimating correlation parameters \( \rho \) (Hamza et al., 2008). In fact, the FL method encounters the issues of non-convergence and a singular covariance matrix when used to analyse the motivating study on metastases. Specifically, for the subgroup of the PET–CT test with a total of...
eight studies, we failed to obtain the maximum likelihood estimates as the likelihood in equation (4) contains nine parameters. For the subgroup of the CT test with a total of 12 studies, the convergence of the FL method heavily depends on the choice of the initial values: the FL method leads to non-convergent results for some default choice of initial values. For the subgroup of the PET test with a total of 29 studies, a ‘poor’ choice of initial value results in a singular covariance matrix. These results are summarized in Table S2 in the on-line supplemental material.

As suggested in equation (3), the commonly used measures of diagnostic tests are functions of $(\beta_0, \beta_1, \beta_2)$ only and do not involve the correlation parameters. Therefore, we propose an alternative inference procedure that focuses on $(\beta_0, \beta_1, \beta_2)$ without inferring $\rho$. The key of the procedure proposed is the factorization of the multinomial likelihood function as a product of three independent binomial likelihoods. Specifically, the likelihood function based on a cohort study for given $(\pi_i, Se_i, Sp_i)$ can be factored as

$$\text{multinomial}\{n_{i11}, n_{i10}, n_{i01}, n_{i00}; \pi_i \ Se_i, (1 - \pi_i)(1 - Sp_i), \pi_i(1 - Se_i), (1 - \pi_i) Sp_i\}$$

$$\propto \text{binomial}(n_{i11}|n_i; \pi_i) \text{binomial}(n_{i11}|n_i; Se_i) \text{binomial}(n_{i00}|n_i; Sp_i),$$

where $i = m_1 + 1, \ldots, m$. Given this factorization, we can construct a CL function under a working independence assumption. Mathematically, by letting $\rho_{01} = \rho_{02} = \rho_{12} = 0$ in equation (4), we obtain the CL function

$$\log \{L_c(\theta_0, \theta_1, \theta_2)\} = \log \{L_0(\theta_0)\} + \log \{L_1(\theta_1)\} + \log \{L_2(\theta_2)\}$$  \hspace{1cm} (5)

where

$$\log \{L_0(\theta_0)\} = \sum_{i=m_1+1}^{m} \log \{\Pr(n_{i11}|n_i; \theta_0)\} = \sum_{i=m_1+1}^{m} \log \left\{ \int \text{binomial}(n_{i11}|n_i, \pi_i) \phi(\pi_i; \theta_0) d\pi_i \right\},$$

$$\log \{L_1(\theta_1)\} = \sum_{i=1}^{m} \log \{\Pr(n_{i11}|n_i; \theta_1)\} = \sum_{i=1}^{m} \log \left\{ \int \text{binomial}(n_{i11}|n_i, Se_i) \phi(Se_i; \theta_1) dSe_i \right\},$$

$$\log \{L_2(\theta_2)\} = \sum_{i=1}^{m} \log \{\Pr(n_{i00}|n_i; \theta_2)\} = \sum_{i=1}^{m} \log \left\{ \int \text{binomial}(n_{i00}|n_i, Sp_i) \phi(Sp_i; \theta_2) dSp_i \right\},$$

and $\phi(\cdot; \theta_j)$ is the univariate logit normal distribution with mean $\beta_j$ and variance $\tau_j^2$ on the logit scale, and is indexed by $\theta_j$ ($j = 0, 1, 2$).

Since each component of the CL function, $\{L_j(\theta_j)\}$ ($j = 0, 1, 2$), is a true log-marginal-likelihood, the score equation of CL is unbiased. Consequently, the estimator $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$, defined as a solution of the score equation, is consistent and asymptotically normal. However, the conventional covariance matrix estimator $I_c(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)^{-1}$, where

$$I_c(\theta_0, \theta_1, \theta_2) = -\frac{\partial^2 \log \{L_c(\theta_0, \theta_1, \theta_2)\}}{\partial(\theta_0, \theta_1, \theta_2)^2},$$

is no longer valid because $E\{I_c(\theta_0, \theta_1, \theta_2)\}$ is not the covariance matrix of $\partial L_c(\theta_0, \theta_1, \theta_2)/\partial(\theta_0, \theta_1, \theta_2)$ in the presence of correlations between $(\pi_i, Se_i, Sp_i)$.

Assume that $m_2/m \rightarrow r > 0$ as $m \rightarrow \infty$. As shown in section 1 of the on-line supplementary material, the estimator $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$ is asymptotically normal with mean 0 and symmetric covariance matrix.
\[
\Sigma = \left( \begin{array}{cccc}
    r^{-1}I_{00}^{-1} & r^{-1/2}I_{00}^{-1/2}I_{01}I_{11}^{-1} & r^{-1/2}I_{00}^{-1/2}I_{02}I_{22}^{-1} \\
    I_{11}^{-1} & I_{11}^{-1} & I_{12}I_{22}^{-1} \\
    I_{12}^{-1} & I_{12}^{-1} \\
\end{array} \right),
\]

where

\[
I_{00} = E \left[ -\frac{1}{m_2} \frac{\partial^2 \log \{ L_0(\theta_0) \} }{\partial \theta_0^2} \right],
\]
\[
I_{jj} = E \left[ -\frac{1}{m} \frac{\partial^2 \log \{ L_j(\theta_j) \} }{\partial \theta_j^2} \right],
\]
\[
I_{12} = E \left[ \frac{1}{m} \left( \frac{\partial \log \{ L_1(\theta_1) \} }{\partial \theta_1} \right) \left( \frac{\partial \log \{ L_2(\theta_2) \} }{\partial \theta_2} \right)^T \right]
\]
and

\[
I_{0j} = E \left[ \frac{1}{m_2} \left( \frac{\partial \log \{ L_0(\theta_0) \} }{\partial \theta_0} \right) \left( \frac{\partial \log \{ L_j(\theta_j) \} }{\partial \theta_j} \right)^T \right]
\]

for \( j = 1, 2 \). In practice, the asymptotic covariance matrix \( \Sigma \) can be consistently estimated by its empirical counterpart \( \tilde{\Sigma} \) as follows:

\[
\tilde{\Sigma} = \left( \begin{array}{cccc}
    \tilde{r}^{-1} \tilde{I}_{00}^{-1} & \tilde{r}^{-1/2} \tilde{I}_{00}^{-1/2} \tilde{I}_{01} \tilde{I}_{11}^{-1} & \tilde{r}^{-1/2} \tilde{I}_{00}^{-1/2} \tilde{I}_{02} \tilde{I}_{22}^{-1} \\
    \tilde{I}_{11}^{-1} & \tilde{I}_{11}^{-1} & \tilde{I}_{12} \tilde{I}_{22}^{-1} \\
    \tilde{I}_{12}^{-1} & \tilde{I}_{12}^{-1} \\
\end{array} \right),
\]

where \( \tilde{r} = m_2/m \),

\[
\tilde{I}_{00} = -\frac{1}{m_2} \sum_{i=m_1+1}^m \frac{\partial^2 \log \{ \Pr(n_{i1} | n_{i}; \tilde{\theta}_0) \} }{\partial \theta_0^2},
\]
\[
\tilde{I}_{11} = -\frac{1}{m} \sum_{i=1}^m \frac{\partial^2 \log \{ \Pr(n_{i11} | n_{i1}; \tilde{\theta}_1) \} }{\partial \theta_1^2},
\]
\[
\tilde{I}_{22} = -\frac{1}{m} \sum_{i=1}^m \frac{\partial^2 \log \{ \Pr(n_{i00} | n_{i0}; \tilde{\theta}_2) \} }{\partial \theta_2^2},
\]
\[
\tilde{I}_{01} = \frac{1}{m_2} \sum_{i=m_1+1}^m \left( \frac{\partial \log \{ \Pr(n_{i1} | n_{i}; \tilde{\theta}_0) \} }{\partial \theta_0} \right) \left( \frac{\partial \log \{ \Pr(n_{i11} | n_{i1}; \tilde{\theta}_1) \} }{\partial \theta_1} \right)^T,
\]
\[
\tilde{I}_{02} = \frac{1}{m_2} \sum_{i=m_1+1}^m \left( \frac{\partial \log \{ \Pr(n_{i1} | n_{i}; \tilde{\theta}_0) \} }{\partial \theta_0} \right) \left( \frac{\partial \log \{ \Pr(n_{i00} | n_{i0}; \tilde{\theta}_2) \} }{\partial \theta_2} \right)^T
\]
and

\[
\tilde{I}_{12} = \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \log \{ \Pr(n_{i11} | n_{i1}; \tilde{\theta}_1) \} }{\partial \theta_1} \right) \left( \frac{\partial \log \{ \Pr(n_{i00} | n_{i0}; \tilde{\theta}_2) \} }{\partial \theta_2} \right)^T.
\]
The CL method reduces the computationally demanding three-dimensional integrals in the full likelihood to computationally much simpler one-dimensional integrals. More importantly, the non-convergence problem of the FL method is alleviated since no correlation parameter (i.e. $\rho_{01}, \rho_{02}$ or $\rho_{12}$) is involved in the CL. The maximum CL estimate $\hat{\beta}_j$ can be obtained by conducting a univariate meta-analysis with a random-effects model, which is available in most statistical software. The covariance matrix of $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$ can be easily calculated by using the above formulae, which involve only one-dimensional integrals. Note that the off-diagonal matrices in $\Sigma$ properly account for the covariance between the estimated overall disease prevalence and diagnostic test sensitivity and specificity, which is not possible if investigators conduct meta-analysis by univariate meta-analysis. We consider the CL method as a method between multivariate and univariate meta-analyses, inheriting the ability of multivariate meta-analysis to infer functions of overall parameters such as PPV and NPV (i.e. functions of $\beta_0, \beta_1$ and $\beta_2$) but not suffering from its limitations.

The asymptotic results of $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$ can be used to construct approximate Wald-type confidence intervals or regions for diagnostic measures of interest. Alternatively, CL ratio-based inference is available. In general, the CL ratio test statistic converges to a non-standard asymptotic distribution as a weighted sum of independent $\chi^2$-distributions, which can be derived as a special case of results on misspecified likelihoods (Kent, 1982; White, 1996). Several adjustments of CL have been proposed to have an approximate or asymptotic $\chi^2$-distribution (Geys et al., 1999; Rotnitzky and Jewell, 1990; Satterthwaite, 1946; Lindsay et al., 2000; Chandler and Bate, 2007; Pace et al., 2011).

3. Simulation study

To evaluate and compare the finite sample performance of the FL and CL methods, we conduct simulation studies. The data are generated from a two-stage procedure, as specified by equations (1) and (2). Two settings of covariates are considered. In the first setting, there is no study level covariate except the intercept, i.e. $W_i = X_i = Z_i = 1$. In the second setting, we consider two covariates: a binary covariate (e.g. 1 for regional cancer and 0 for distant cancer) and a continuous covariate sampled from a uniform distribution (e.g. quality assessment of diagnostic accuracy studies score with range of 1–14). We consider a moderate size meta-analysis with $m = 30$ studies and a relatively large meta-analysis with $m = 50$ studies. We assume equal numbers of case–control and cohort studies. The numbers of subjects in each study (i.e. $(n_{i0}, n_{i1})$ in case–control studies and $n_i$ in cohort studies) are randomly drawn from the studies on metastases that were described in Section 1. Specifically, in the case-control study, the range of the number of subjects per study is 10–100 for patients with metastasis, and 10–124 for patients without metastasis. In the cohort studies, the number of subjects per study ranges from 20 to 220. For the setting without study level covariate, two configurations of overall disease prevalence and test sensitivity and specificity are considered, namely $\beta_0 = \text{logit}(0.2), \beta_1 = \text{logit}(0.6)$ or $\beta_1 = \text{logit}(0.9)$ and $\beta_2 = \text{logit}(0.9)$. For the setting with study level covariates, the values of regression coefficients are set at the estimates from fitting the model on the studies on metastases. We let the between-study variances in disease prevalence and test sensitivity and specificity be $\tau_0^2 = \tau_1^2 = \tau_2^2 = 1$. To evaluate the effect of the correlation structure on the inference, we let the correlation parameters $(\rho_{01}, \rho_{02}, \rho_{12})$ take values of $(0, 0, 0), (0, 0, -0.6), (0.2, -0.2, -0.6), (0.6, -0.6, -0.6)$ or $(0.8, -0.8, -0.8)$ to represent different levels of correlation between disease prevalence and test sensitivity and specificity (on the logit scale). We also consider the heterogeneity in correlation where the correlation parameters $(\rho_{01}, \rho_{02}, \rho_{12})$ take the values $(0, 0, -0.6)$ in half of the studies and take different values $(0.6, -0.6, -0.6)$ in the remaining half. Under this setting, the likelihood of the
FL method is misspecified, whereas the likelihood of the CL method is not because the CL method does not assume homogeneous correlation across studies. For each simulation setting, we generate 1000 samples. The samples are simulated in R (R Core Team (2014), version 2.14.1). The CL method is implemented in R by using the glmmML package (Broström and Holmberg, 2011). The FL method is implemented in SAS where the adaptive Gaussian quadrature method in the NLMIXED procedure is used to maximize the FL.

Fig. 1 summarizes the empirical bias and the coverage of the 95% confidence intervals of estimates from the FL and CL methods based on 1000 samples when the number of studies is 30. The parameters of interest are the overall disease prevalence $\text{Prev}$, test sensitivity $\text{Se}$ and specificity $\text{Sp}$, and positive and negative predicted values $\text{PPV}$ and $\text{NPV}$. We note that $\text{PPV}$ and $\text{NPV}$ are functions of parameters $(\beta_0, \beta_1, \beta_2)$ as described by equation (3). The delta method was used to derive their standard errors. From Figs 1(a) and 1(b), the FL method gives approximately unbiased estimates and its coverage is close to the nominal level when the correlations are zero, i.e. $(\rho_{01}, \rho_{02}, \rho_{12}) = (0, 0, 0)$. When the correlations become stronger, the estimates from the FL method are still approximately unbiased, but the coverage of the Wald-based confidence intervals deteriorate (range of coverage, 76.3–95.1%). Furthermore, when the correlation structures are heterogeneous across studies (denoted as ‘heterogeneous’), the bias of the FL method becomes larger and its coverage is below 70%. In contrast, the CL method provides approximately unbiased estimates and confidence intervals with better coverage under all the correlation structures that were considered (range of coverage, 85.3–95.3%), including the scenario with heterogeneous correlation structures. We note that the non-convergence rate (i.e. the number of iterations reaches the default number of 200 iterations whereas the relative gradient convergence criterion of less than $10^{-10}$ is not satisfied) for the FL method increases as the degree of correlation increases, and varies from 10.9% to a substantial proportion of 47.6%, whereas the non-convergence rate for the CL method is less than 1.5% under all the settings that were considered. We also calculate the relative efficiency RE of the CL method, defined as the square of the empirical standard error of the estimates from the CL method, divided by that of the FL method. The range of RE under all correlation structures except the heterogeneous structure is 76.6–122.2%. The loss of efficiency is expected because the FL method is asymptotically the most efficient method, whereas the gain in efficiency can be explained by the advantage of not estimating the correlation parameters in the CL method. There is an 82–109% gain in efficiency in the CL method under the heterogeneous correlation structure setting due to the better fit of the CL method compared with the FL method. We also conducted simulations for scenarios with $m = 8$ and $m = 50$ studies, and similar findings are obtained in that the CL method has better coverage, avoids the non-convergence problem and is robust to the heterogeneous correlation structures. We note that when the number of studies is small (e.g. $m = 8$), although the CL method has substantially better coverage than the FL method (range of coverage, 77.5–89.4% versus 29.3–88.7%), bootstrap standard errors from the CL method should be used for more satisfactory coverage. The detailed simulation results are summarized in Tables S3–S8 in the on-line supplementary material.

Fig. 2 summarizes the simulation results when study level covariates are available and the number of studies is 30. In this case, the regression coefficients are the parameters of interest. The CL method has slightly larger bias but much better coverage than the FL method. The range of the non-convergence rate of the FL method is 10.1–12.7%, whereas the non-convergence rate of the CL method is less than 1.5%. The range of RE of the CL method is 55.9–116.6%. There is a substantial loss of efficiency for the CL method when the correlations are high (i.e. $\rho_{01} = 0.8$, $\rho_{02} = -0.8$ and $\rho_{12} = -0.8$). However, the corresponding coverage of $\beta_{00}$ is 90.6% for the CL method and 78.9% for the FL method when RE is 56.0%. This suggests that choosing
Fig. 1. Bias and coverage for estimated disease prevalence (○), sensitivity (□), specificity (◊), PPV (△) and NPV (⋆) from (a), (b) the FL and (c), (d) CL methods (the true overall disease prevalence is 0.2, the sensitivity is 0.9 and the specificity is 0.9; the data are generated from a bivariate GLMM (for the case–control studies) and a trivariate GLMM (for the cohort studies); the results are summarized from 1000 simulations; the x-axis represents the various settings of pairwise correlations among study-specific prevalence, sensitivity and specificity (on the logit scale)): (a), (c) bias; (b), (d) coverage
Fig. 2. Bias and coverage for estimated meta-regression parameters in equation (2) from (a)–(f) the FL and (g)–(l) the CL methods (the true values of the regression parameters are $(\hat{\beta}_{00}, \hat{\beta}_{01}, \hat{\beta}_{02}) = (0.173, -1.295, 0)$, $(\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}) = (1.712, -1.266, 0)$ and $(\hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}) = (1.912, 1.263, 0)$; the data are generated from a bi-variate GLMM (for the case–control studies) and a trivariate GLMM (for the cohort studies); the results are summarized from 1000 simulations; the $x$-axis represents for the various settings of pairwise correlations among study-specific prevalence, sensitivity and specificity (on the logit scale): (a)–(c), (g)–(i) bias; (d)–(f), (j)–(l) coverage.

The FL method over the CL method in this setting to achieve better efficiency comes at the cost of a much lower coverage. A similar finding is obtained when the number of studies is 50; these results are summarized in Tables S9 and S10 in the on-line supplementary material.

To investigate the robustness of both methods to misspecifications of the model, we generate study-specific prevalence, sensitivity and specificity from a trivariate $t$-distribution with 4 degrees of freedom. This setting mimics the situation in which the distributions have heavier tails than those of the normal distributions. Under this setting, both the likelihood of the FL method and the likelihood of the CL method are misspecified. Fig. 3 summarizes the bias and coverage from 1000 simulations for various correlation structures when the number of studies is 30. The simulation results suggest that, despite the misspecification, the bias of both methods is in a reasonable range. The only exception is that the FL method has relatively large biases under
Combining Case–Control and Cohort Studies in Systematic Reviews

Fig. 3. Bias and coverage for estimated disease prevalence (○), sensitivity (□), specificity (△), PPV (△) and NPV (⋆) from (a), (b) the FL and (c), (d) CL methods (the true overall disease prevalence is 0.2, the sensitivity is 0.9 and the specificity is 0.9; the data are generated from a bivariate $t$-distribution (for the case–control studies) and a trivariate $t$-distribution (for the cohort studies); the results are summarized from 1000 simulations; the $x$-axis represents the various settings of pairwise correlations between study-specific prevalence, sensitivity and specificity (on the logit scale)): (a), (c) bias; (b), (d) coverage.
heterogeneous correlation structures. The coverage of the FL method is close to the nominal level only when the correlations are small, and it deteriorates quickly as the correlation increases. In contrast, the CL method has better coverage in all the settings that were considered. Similar findings are obtained when the number of studies is 8 or 50. These results are summarized in Tables S11–S16 in the online supplementary material.

In summary, both FL and CL methods perform well when the number of studies is relatively large and the correlations are relatively small, and the CL method outperforms the FL method when the number of studies is relatively small, or when the correlations are relatively large. In addition, the CL method has certain robustness to heterogeneous correlation structures, and model misspecifications. The CL method maintains relatively high efficiency except that the correlations are exceptionally high, which is consistent with previous findings in settings of longitudinal data analyses (Liang and Zeger, 1986; McDonald, 1993; Sutradhar and Das, 1999; Henderson and Shimakura, 2003). Considering these performances of the CL method, and its computational advantages, we recommend the use of the CL method for practical investigators.

4. Systematic review of modern diagnostic imaging modalities for surveillance of melanoma patients

We apply the model proposed and the CL method to a systematic review of published studies which examined diagnostic modality characteristics for melanoma. This systematic review contains 98 studies that had obtained data from 10528 patients with melanoma between January 1st, 1990, and June 30th, 2009 (Xing et al., 2011). As mentioned, the available studies include 41 case–control studies and 57 cohort studies. To combine the case–control and cohort studies effectively, we fit the model that is described by equations (1) and (2). A sequence of nested models is fitted where the smallest model (referred to as the baseline model) includes a variable for the stage of cancer (i.e. 1 for regional and 0 for distant), and three dummy variables for types of imaging modalities with PET–CT as the reference group. Larger models were considered by including interaction terms between type of cancer and imaging modalities. For the CL method, modifications of Akaike’s information criterion (AIC) and the Bayesian information criterion (BIC) have been proposed in the literature (Varin and Vidoni, 2005; Gao and Song, 2010). Specifically, the CL version of the AIC is defined as (Varin and Vidoni, 2005)

$$\text{CL-AIC} = -2 \log(L_c) + 2d^*_s$$

where $d^*_s = \text{tr}\{\hat{J}(\hat{H})^{-1}\}$, and ‘s’ denotes a candidate marginal submodel s, $\hat{J}$ is the estimated covariance matrix of $\partial L_c(\theta_0, \theta_1, \theta_2)/\partial(\theta_0, \theta_1, \theta_2)$ evaluated at $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$ and $\hat{H}$ is $-\partial^2 \log\{L_c(\theta_0, \theta_1, \theta_2)\}/\partial(\theta_0, \theta_1, \theta_2)^2$ evaluated at $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$. In section 2 of the online supplementary materials, we show that for our model $d^*_s$ converges with increasing m to the number of parameters in the model. The CL version of the BIC is defined as (Gao and Song, 2010)

$$\text{CL-BIC} = -2 \log(L_c) + d^*_s \log(m) + 2\gamma d^*_s \log(P)$$

where $P$ is the number of model parameters and $\gamma$ is a tuning parameter and is taken as 0 when $P$ is relatively small compared with the number of studies as suggested in Gao and Song (2010). The results of fitting the sequence of nested models are summarized in Table 2. Both CL versions of the AIC and BIC suggest the use of the baseline model with 18 model parameters.

To investigate the model assumptions, we start with checking the normality assumption on the logit prevalence, sensitivity and specificity. QQ-plots are provided in Fig. S3 of the online supplementary material. A test of normality is conducted by the Shapiro–Wilk test (Shapiro
Table 2. Model selection using CL-AIC and CL-BIC when analysing the data in Xing et al. (2011)†

<table>
<thead>
<tr>
<th>Model</th>
<th>( d^2 )</th>
<th>(-2 \log(\text{CL}))</th>
<th>CL-AIC</th>
<th>CL-BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>18</td>
<td>1537</td>
<td>1573</td>
<td>1620</td>
</tr>
<tr>
<td>+ ( I(\text{Regional}) + I(\text{US}) )</td>
<td>21</td>
<td>1534</td>
<td>1576</td>
<td>1632</td>
</tr>
<tr>
<td>+ ( I(\text{Regional}) + I(\text{CT}) )</td>
<td>21</td>
<td>1535</td>
<td>1577</td>
<td>1632</td>
</tr>
<tr>
<td>+ ( I(\text{Regional}) + I(\text{PET}) )</td>
<td>21</td>
<td>1534</td>
<td>1576</td>
<td>1632</td>
</tr>
<tr>
<td>+ ( I(\text{Regional}) + I(\text{US}) + I(\text{Regional}) + I(\text{CT}) )</td>
<td>24</td>
<td>1532</td>
<td>1580</td>
<td>1643</td>
</tr>
<tr>
<td>+ ( I(\text{Regional}) + I(\text{US}) + I(\text{Regional}) + I(\text{PET}) )</td>
<td>24</td>
<td>1529</td>
<td>1577</td>
<td>1641</td>
</tr>
<tr>
<td>+ ( I(\text{Regional}) + I(\text{CT}) + I(\text{Regional}) + I(\text{PET}) )</td>
<td>24</td>
<td>1533</td>
<td>1581</td>
<td>1644</td>
</tr>
<tr>
<td>+ ( I(\text{Regional}) + I(\text{US}) + I(\text{Regional}) + I(\text{CT}) )</td>
<td>27</td>
<td>1525</td>
<td>1579</td>
<td>1650</td>
</tr>
</tbody>
</table>

†US, ultrasonography; baseline, meta-analysis model with study level covariates of \( I(\text{Regional}) + I(\text{US}) + I(\text{CT}) + I(\text{PET}) \).

and Wilk, 1965) for each subgroup and the \( p \)-values are all greater than 0.05, suggesting that the normality assumption is appropriate. The sequence of models in Table 2 implicitly assumes equal variance in logit prevalence, sensitivity and specificity for different subgroups. To check such an assumption, we apply Bartlett's test (Bartlett, 1937) for the homogeneity in variances across subgroups. The test suggests that the homogeneity assumption is appropriate for logit prevalence and specificity, but not for logit sensitivity (\( p < 0.001 \)). To study the sensitivity of the results from the baseline model (as recommended by both CL-AIC and CL-BIC) to the equal variance assumption, we conduct an alternative analysis within each of the subgroups (stratified by stage of cancer, and type of imaging modality). We found that the results from the subgroup analyses are generally similar to those from the baseline model. There are only two case–control studies and one cohort study in the CT subgroup for regional cancer. In this subgroup, the CL method cannot be applied since the model for the prevalence of metastasis contains two parameters and the estimation requires at least two cohort studies. Instead, the bivariate GLMM was fitted to obtain estimates of diagnostic sensitivity and specificity, but not metastasis prevalence, PPV or NPV.

Fig. 4 presents the results from the subgroup analyses with the CL method on the overall prevalence of metastasis, diagnostic sensitivity and specificity, PPV and NPV, and the associated 95% confidence intervals for the four diagnostic imaging modalities. The confidence intervals are symmetric around the estimates on the logit scale and asymmetric on the original scale. The results from fitting the bivariate GLMM are displayed as the broken lines in Fig. 4 for comparison. In general, our results are consistent with those from the bivariate GLMM, with respect to sensitivity and specificity. Here we highlight selected results of diagnostic sensitivities and specificities, and PPV and NPV. For the surveillance of regional lymph node metastasis, ultrasonography has the highest sensitivity (68%; 95% confidence interval CI = 44–85%) and specificity (98%; 95% CI = 96–99%) among all four imaging modalities. In contrast, patients diagnosed by PET–CT or PET have higher estimated prevalences of metastasis and hence higher estimated PPV, compared with patients diagnosed by ultrasonography. For the surveillance of distant lymph node metastasis, PET–CT has the highest sensitivity (87%; 95% CI = 75–94%), specificity (94%; 95% CI = 88–97%), PPV (93%; 95% CI = 83–97%) and NPV (87%; 95% CI = 83–91%). There is a significant heterogeneity in prevalence of metastasis across different imaging modalities and stage of metastasis (regional versus distant). These differences are potentially meaningful in practice. Patients with distant metastasis have higher PPV than patients with
metastasis that is confined to regional lymph nodes. The results of this systematic review of imaging modalities used to stage melanoma suggest that ultrasonography is a more accurate imaging modality for diagnosing lymph node involvement and PET–CT is the preferred imaging modality to diagnose distant metastasis. In addition, owing to the low prevalence of metastasis among patients for whom lymph node metastasis is diagnosed by ultrasonography, a positive test result from ultrasonography yields the lowest PPV among all imaging modalities.

As we know, a univariate summary measure of diagnostic tests may not be sufficient, and the use of bivariate summary measures is preferred when describing diagnostic tests such as (sensitivity, specificity) or (PPV, NPV). Additionally, estimates of these bivariate summary measures are often correlated. Therefore, separate confidence intervals that do not account for such a correlation may be misleading (Harbord et al., 2007), and confidence regions should be used. Fig. 5 shows the summary points and 95% confidence regions for sensitivity versus 1 minus specificity (Fig. 5(a)), PPV versus NPV (Fig. 5(b)), sensitivity and specificity versus metastasis prevalence (Figs 5(c) and 5(d)) and predictive values versus metastasis prevalence (Figs 5(e) and 5(f)) without stratification on stages of metastasis (i.e. regional or distant). These regions are calculated as Wald-based confidence regions and are not elliptical because they are on the original scale. The elliptical confidence regions on the logit scale are displayed in Fig. S4 in the on-line supplementary material. Specifically, following Douglas (1993), the parametric representation of the boundary of the elliptical Wald-type confidence region for sensitivity and specificity (on the logit scale) is
Fig. 5. Summary points and 95% confidence regions of (a) sensitivity versus 1 – specificity, (b) PPV versus NPV, (c) sensitivity versus prevalence of metastasis, (d) specificity versus prevalence of metastasis, (e) PPV versus prevalence of metastasis and (f) NPV versus prevalence of metastasis for four diagnostic imaging modalities: •, summary point; ———, boundary of the 95% confidence region for the summary point.
Fig. 6. Overall PPV and NPV plots for a given prevalence based on the meta-analysis without study level covariate by using the CL method (—, estimate; ——, 95% confidence interval; i, estimated overall prevalence): (a) ultrasonography; (b) CT; (c) PET; (d) PET-CT

$S_1 = \hat{S}_1 + s_{S_1} \sqrt{(2f_{2,n-2;\alpha}) \cos(\phi)}$

and

$C_1 = \hat{C}_1 + s_{C_1} \sqrt{(2f_{2,n-2;\alpha}) \cos\{\phi + \cos^{-1}(r)\}},$

where $s_{S_1}$ and $s_{C_1}$ are the estimated standard errors of $\hat{S}_1$ and $\hat{C}_1$, $r$ is the estimate of their correlation, $\phi$ runs from 0 to $2\pi$ and $f_{2,n-2;\alpha}$ is the upper 100\% point of the $F$-distribution with degrees of freedom 2 and $n-2$, and $n$ is the number of studies. For joint confidence regions of PPV and NPV (or pairs of other measures), the delta method was used to obtain the covariance matrix.

As suggested by Fig. 5(a), there is more variation in sensitivity than in specificity. This is because the number of true positive results is much less than the number of true negative results. The PPV and NPV tend to have similar variations as seen in Fig. 5(b). Figs 5(c) and 5(d) suggest that the confidence region covers a larger range of prevalence than specificity, which suggests more variation in estimated prevalence than specificity. Such a finding is consistent with that from Fig. 4(a), where significant heterogeneity in the estimated prevalence of metastasis is found.
As seen in Figs 5(d) and 5(e), PPV and NPV tend to have similar variation compared with the prevalence. In summary, given the wide ranges of those confidence regions, it suggests that more studies are needed to increase the precision of those estimates, and to reach definitive conclusions comparing those imaging modalities.

Fig. 6 shows the estimated PPV and NPV with their pointwise 95% confidence intervals based on the overall estimates of sensitivity and specificity for each of imaging modalities. Fig. 6 is particularly useful for clinicians who want to obtain PPV and NPV for a different cohort of patients under investigation. The vertical broken lines indicate the estimated prevalence of metastasis for patients diagnosed by the imaging modality. For example, the estimated prevalence for patients diagnosed by ultrasonography is 15%, and the estimated overall PPV and NPV are 85% (95% CI = 80–88%) and 95% (95% CI = 89–98%) respectively. In contrast, the estimated overall PPV and NPV for patients diagnosed by CT are 79% (95% CI = 76–82%) and 78% (95% CI = 71–83%) with estimated prevalence being 42%. This suggests that ultrasonography is more useful than CT in ruling out disease and detecting disease for patients diagnosed by the corresponding imaging modality. We note that we did not stratify by stage of cancer in this analysis as the number of studies stratified by both stage of cancer and type of imaging modality is very limited.

5. Discussion

Multivariate meta-analysis has gained in popularity recently, especially in systematic reviews of diagnostic tests (Jackson et al., 2011). In this paper, we proposed a hybrid multivariate random-effects model for study of diagnostic test accuracy. There are two major advantages of multivariate meta-analysis over univariate meta-analysis. First, unlike univariate meta-analysis, multivariate meta-analysis can provide valid inference on functions of overall population parameters, such as PPV and NPV. Secondly, by jointly modelling the study-specific effects, multivariate meta-analysis is expected to have more efficiency than univariate meta-analysis in terms of parameter estimation. In contrast, multivariate meta-analysis may suffer the non-convergence problem and computational difficulties, especially when the number of studies is relatively small, which is common in practical meta-analysis.

In this paper, we propose the CL inference procedure, which can be thought of as a procedure between multivariate and univariate meta-analyses, inheriting the ability of multivariate meta-analysis to infer functions of overall parameters while not suffering from their limitations. Through simulation studies, we find that CL inference does not suffer a severe loss of efficiency except in situations with exceptionally high correlations. CL inference is also more robust than the FL inference to model misspecifications. Therefore, the CL method can serve as a useful alternative in multivariate meta-analysis of diagnostic tests.

Acknowledgements

We are grateful to the Joint Editor, the Associate Editor and the reviewer, whose suggestions have greatly improved the presentation of this work. Yong Chen was supported by grant R03HS022900 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. Jing Ning was partially supported by a start-up fund from the University of Texas MD Anderson Cancer Center. Haitao Chu was supported in part by the US National Institute of Allergy and Infectious Diseases grant AI103012, AHRQ R03HS020666, National Cancer Institute grant 1P01CA142538 and National Cancer Institute grant 2P30CA077598.
References


**Supporting information**

Additional ‘supporting information’ may be found in the on-line version of this article:

“Supplementary Materials for “A hybrid model for combining case-control and cohort studies in systematic reviews of diagnostic tests”.”