6-22-2007

RANDOM EFFECTS MODELS IN A META-ANALYSIS OF THE ACCURACY OF DIAGNOSTIC TESTS WITHIN A GOLD STANDARD IN THE PRESENCE OF MISSING DATA

Haitao Chu
Department of Biostatistics and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill,
hchu@jhsph.edu

Sining Chen
Johns Hopkins Bloomberg School of Public Health, Department of Environmental Health Sciences

Thomas A. Louis
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Random Effects Models in a Meta-Analysis of the Accuracy of Diagnostic Tests without a Gold Standard in the Presence of Missing Data

Haitao Chu[*1], Sining Chen[2], Thomas A. Louis[3]

[1]Department of Biostatistics and Lineberger Comprehensive Cancer Center

The University of North Carolina at Chapel Hill

Chapel Hill, NC 27599

[2]Department of Environment Health

The Johns Hopkins Bloomberg School of Public Health

Baltimore, MD 21205 USA

[3]Department of Biostatistics

The Johns Hopkins Bloomberg School of Public Health

Baltimore, MD 21205 USA

[*] Corresponding Author

Email: hchu@bios.unc.edu

Fax: 919-966-4244 Phone: 919-966-5269

21 June 2007

Acknowledgement: The authors are very grateful to Dr. Giovanni Parmigiani for helpful comments.

Random Effects Models in a Meta-Analysis of the Accuracy of Diagnostic Tests without a Gold Standard in the Presence of Missing Data

Summary

In evaluating the accuracy of diagnosis tests, it is common to apply two imperfect tests jointly or sequentially to a study population. In a recent meta-analysis of the accuracy of microsatellite instability testing (MSI) and traditional mutation analysis (MUT) in predicting germline mutations of the mismatch repair (MMR) genes, a Bayesian approach (Chen, Watson, and Parmigiani 2005) was proposed to handle missing data resulting from partial testing and the lack of a gold standard. In this paper, we demonstrate an improved estimation of the sensitivities and specificities of MSI and MUT by using a nonlinear mixed model and a Bayesian hierarchical model, both of which account for the heterogeneity across studies through study-specific random effects. The methods can be used to estimate the accuracy of two imperfect diagnostic tests in other meta-analyses when the prevalence of disease, the sensitivities and/or the specificities of diagnostic tests are heterogeneous among studies. Furthermore, simulation studies have demonstrated the importance of carefully selecting appropriate random effects on the estimation of diagnostic accuracy measurements in this scenario.

Key Words: diagnostic test; generalized linear mixed model; missing data; gold standard; meta-analysis; Bayesian hierarchical model.

Running Title: Multivariate meta-analysis of diagnostic tests.

1. Introduction

The performance of a binary diagnostic test is usually represented by sensitivity (Se) and specificity (Sp). Sensitivity is also referred to as the true positive fraction, defined as the probability of testing positive given diseased person. Specificity is also known as the true negative fraction, defined as the probability of test negative given non-diseased person (Zhou, Obuchowski, and McClish 2002; Pepe 2003). The true disease status is usually measured by a "gold standard" test. However, a definitive "gold standard" may not be available for some diseases. For example, due to the limitation of laboratory techniques, it is difficult to diagnose Lynch syndrome with certainty (Lynch and de la Chapelle 1999), caused by a deleterious germline mutation in one of the mismatch repair (MMR) genes, mainly MSH2 and MLH1.

When a "gold standard" is not readily available, it is common to apply two or more imperfect screening or diagnostic tests to improve accuracy. There is a considerable literature discussing the challenges and approaches to assess the performance of diagnostic tests from a single population (Gart and Buck 1966; Joseph, Gyorkos, and Coupal 1995; Andersen 1997; Johnson, Gastwirth, and Pearson 2001). Under the assumption that the two imperfect tests are conditionally independent given the latent true disease status, the challenge is to estimate five parameters (i.e., prevalence, the two sensitivities and the two specificities) from only three independent cells in a two by two table. Since the model is over-parameterized and not identifiable, even Bayesian approaches — which can take advantage of prior knowledge about the accuracy of the tests and the disease prevalence — do not generally converge to the "true" values as the sample size increases (Johnson, Gastwirth, and Pearson 2001).

To overcome the identifiability problem, sampling from a second population with a different prevalence was suggested (Hui and Walter 1980). Assuming that the tests have the same accuracy measures in both populations, there are six conditionally independent cells which provide enough degrees of freedom to estimate the six parameters (including two prevalences, two sensitivities and two specificities). In a recent meta-analysis of seventeen studies to evaluate the accuracy of microsatellite instability testing (MSI) and traditional mutation analysis (MUT) in predicting germline mutations of mismatch repair (MMR) genes, a Bayesian approach was proposed to handle missing data resulting from partial testing (Chen, Watson, and Parmigiani 2005). However, the meta-analysis assumed that the sensitivity of both tests does not differ from study to study, and after categorizing the studies into a high risk and a low risk groups, the prevalence is homogeneous within each group.

In this article, we relax the above assumptions. We present improved methods for estimating the sensitivities and specificities of two imperfect tests based on meta-analysis using a nonlinear mixed effects model, which take into account heterogeneity across studies through study-specific random effects. We reanalyzed the data published in Chen et al. (2005) using the improved methods.

2. Study Background

We introduce the scientific question that Chen et al. (2005) and we address. The DNA mismatch repair (MMR) system repairs the mismatches in the genome that occur during cell duplication. When a person carries a deleterious germline mutation in the MMR genes, mainly MLH1 and MSH2, the impaired mismatch repair mechanism gives rise to the most common hereditary colorectal cancer syndrome, Lynch syndrome (also

known as Hereditary Nonpolyposis Colorectal Cancer, or HNPCC). These Lynch syndrome individuals are also at increased risk of a number of other cancers, most notably endometrial cancer among females.

The diagnosis of Lynch syndrome is synonymous with detecting a mutation in the MMR genes. However, mutation analysis (MUT) of the MMR genes is costly and not always accurate. The main reason is that most mutation analysis techniques fail to detect large genomic deletions and rearrangements, which constitute a significant fraction of all MMR mutations. Because a defective DNA mismatch repair system can also give rise to a tumor phenotype called microsatellite instability (MSI), and MSI testing is relatively inexpensive and believed to have a high negative predictive value, it has become a standard pre-screening procedure for Lynch syndrome (Umar et al. 2004). However, the positive predictive value of MSI is not known as many sporadic cases (i.e., non-cases who are carries) also exhibit MSI. Therefore, understanding the sensitivity and specificity of MSI testing in predicting Lynch syndrome has become a priority in the identification of MMR mutation carriers.

A number of research groups compared MSI results to mutation analysis results in cancer cases. Table 1 presents a list of seventeen studies included in the Meta-analysis in Chen et al (2005). Besides the lack of a gold standard and potential heterogeneity across studies, another challenge is the different patterns of missing data due to partial testing. The most common missing pattern is that due to the perceived high negative predictive value of MSI testing, many studies did not perform mutation analysis on patients whose MSI result is negative.

Fortunately, the assumption that large genomic deletions and rearrangements do not differ from the other mutations in their ability to generate microsatellite instable tumors is biologically reasonable. This gives independence between MSI test result and mutation analysis result conditional on the true mutation status. Furthermore, based on the studies' description of subject ascertainment criteria, they can be categorized into low risk or high risk groups. We will estimate the accuracy of MSI testing based on the above assumptions and observations.

3. Statistical Methods

3.1. Notation and the likelihood function

Let $(P_{i11}, P_{i10}, P_{i01}, P_{i00})$ be the probabilities of MSI and MUT both being positive, MSI positive and MUT negative, MSI negative and MUT positive, MSI and MUT both being negative respectively in study i for $i = 1, \ldots, I$. To describe missing data patterns due to partial testing, the following three categories are involved: (I) MSI and MUT both measured; (II) MSI measured and MUT unmeasured; (III) MSI unmeasured and MUT measured. We define the selection probabilities in the three categories as follows: P_{iA} = Pr (selected to measure MSI only), and P_{iB} = Pr (selected to measure MUT only), from which it follows that the probability of category I is $1 - P_{iA} - P_{iB}$. Table 2 presents a typical data structure for study i when a subset of individuals is only tested by MSI or MUT.

Under the assumption of missing at random for the selection process in the sense of Little and Rubin (Little and Rubin 2002), the likelihood function can be factored into $L(\boldsymbol{\theta}_i, \vartheta_i \mid data) = L(\boldsymbol{\theta}_i \mid data) \times L(\vartheta_i \mid data)$ where $\vartheta_i = (P_{iA}, P_{iB})$ and $\boldsymbol{\theta}_i = (P_{i11}, P_{i10}, P_{i01}, P_{i00})$. Assume the independence of study results conditional on $\boldsymbol{\theta}_i$, the log likelihood for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_I)$ is the summation of the contribution from each study, that is

$$LogL(\boldsymbol{\theta} \mid data) = \sum_i \{ n_{i11}\log(P_{i11}) + n_{i10}\log(P_{i10}) + n_{i01}\log(P_{i01}) + n_{i00}\log(P_{i00}) + n_{i1m}\log(P_{i11}+P_{i10}) + n_{i0m}\log(P_{i01}+P_{i00}) + n_{im1}\log(P_{i11}+P_{i01}) + n_{im0}\log(P_{i10}+P_{i00}) \}$$

Let π_i be the prevalence of true disease, $(Se_{iA}, Se_{iB}, Sp_{iA}, Sp_{iB})$ be the latent sensitivities and specificities for MSI and MUT in study i respectively. Under the assumption of independence for the two testing procedures, we have the following relationship,

$$P_{i11} = \pi_i Se_{iA} Se_{iB} + (1-\pi_i)(1-Sp_{iA})(1-Sp_{iB}),$$

$$P_{i10} = \pi_i Se_{iA}(1-Se_{iB}) + (1-\pi_i)(1-Sp_{iA})Sp_{iB},$$

$$P_{i01} = \pi_i (1-Se_{iA})Se_{iB} + (1-\pi_i)Sp_{iA}(1-Sp_{iB}),$$

$$P_{i00} = \pi_i (1-Se_{iA})(1-Se_{iB}) + (1-\pi_i)Sp_{iA}Sp_{iB}.$$

3.2. Accuracy of diagnostic tests based on a random effects model

To take the potential heterogeneity of the prevalence, sensitivities and specificities across studies into account, in general, we consider a random effects model, specified as follows,

$$\text{logit}(\pi_i \mid \varepsilon_i) = \eta_0 + \eta_1 X_{ij} + \varepsilon_i, \qquad \text{logit}(Sp_{iA} \mid \nu_{iA}) = \beta_A + \nu_{iA},$$

$$\text{logit}(Se_{iA} \mid \mu_{iA}) = \alpha_A + \mu_{iA}, \qquad \text{logit}(Sp_{iB} \mid \nu_{iB}) = \beta_B + \nu_{iB},$$

$$\text{logit}(Se_{iB} \mid \mu_{iB}) = \alpha_B + \mu_{iB}, \qquad (\varepsilon_i, \mu_{iA}, \mu_{iB}, \nu_{iA}, \nu_{iB})' \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\log it(p) = \log(p) - \log(1-p)$, and X_{ij} denotes the risk group status of the j^{th} individual in the i^{th} study with value 1 for being in the high risk group and 0 for being in the lower risk group. The prevalence of disease, π_i, is usually assumed to be independent of sensitivities and specificities of MSI and MUT, thus the variance-covariance matrix $\boldsymbol{\Sigma}$ can be specified as,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{\varepsilon}^2 & 0 & 0 & 0 & 0 \\ & \sigma_{\mu_A}^2 & \rho_A \sigma_{\mu_A} \sigma_{\nu_A} & \rho_{\mu} \sigma_{\mu_A} \sigma_{\mu_B} & \rho_{\mu_A \nu_B} \sigma_{\mu_A} \sigma_{\nu_B} \\ & & \sigma_{\nu_A}^2 & \rho_{\nu_A \mu_B} \sigma_{\nu_A} \sigma_{\mu_B} & \rho_{\nu} \sigma_{\nu_A} \sigma_{\nu_B} \\ & & & \sigma_{\mu_B}^2 & \rho_B \sigma_{\mu_B} \sigma_{\nu_B} \\ & & & & \sigma_{\nu_B}^2 \end{pmatrix}$$

The variance parameters in the diagonal of the variance-covariance matrix $\boldsymbol{\Sigma}$ $\left(\sigma_{\mu_A}^2, \sigma_{\mu_B}^2, \sigma_{\nu_A}^2, \sigma_{\nu_B}^2, \sigma_{\varepsilon}^2\right)$ capture the heterogeneity of sensitivities of MSI and MUT, specificities of MSI and MUT, and the prevalence of disease among studies, respectively. If there is statistical or scientific evidence of homogeneity among studies $\left(\sigma_{\mu_A}^2, \sigma_{\mu_B}^2, \sigma_{\nu_A}^2, \sigma_{\nu_B}^2, \sigma_{\varepsilon}^2 \approx 0\right)$, the corresponding study-specific random effects $\left(\mu_{iA}, \mu_{iB}, \nu_{iA}, \nu_{iB}, \varepsilon_i\right)$ can be dropped from the above model. The parameters $\left(\rho_A, \rho_B, \rho_{\mu}, \rho_{\nu}, \rho_{\nu_A \mu_B}, \rho_{\mu_A \nu_B}\right)$ capture the correlation between Se and Sp of MSI, the correlation between Se and Sp of MUT, the correlation of Se between MSI and MUT, the correlation of Sp between MSI and MUT, the correlation between Se of MSI and Sp of MUT, and the correlation between Sp of MSI and Se of MUT, respectively.

3.3. Parameter estimation and selection of random effects

We adopted two approaches to make inference from the above random effects model. The first is a nonlinear mixed effects model (Davidian and Giltinan 1995; Vonesh and Chinchilli 1997; Molenberghs and Verbeke 2005); and the second is a Bayesian hierarchical model (Carlin and Louis 2000; Gelman et al. 1995). The nonlinear mixed effects model was fitted using PROC NLMIXED in SAS version 9.1 (SAS Institute Inc., Cary, NC). PROC NLMIXED maximizes an approximation to the likelihood integrated over the random effects (Pinheiro and Bates 1995), and the random effects are computed using empirical Bayes estimates. The adaptive Gaussian quadrature approximation and dual

quasi-Newton algorithm optimization techniques in PROC NLMIXED were used to maximize the approximate integrated likelihood. We used PROC NLMIXED built-in ability using the delta method to compute the population estimates of the prevalence for the high and low risk groups, sensitivities and specificities of MSI and MUT on the transformed scale and their confidence intervals based on the normal approximation. In the presence of random effects, those estimates represent the population median estimates. To obtain the population mean estimates, numerical integration over the estimated distributions of random effects can be implemented (Halloran, Preziosi, and Chu 2003).

To avoid over-fitting the data with an excess of random effects (including all of the five random effects in the model), we used a forward selection procedure based on information criteria. The basic idea is to minimize the Kullback-Leibler information (Kullback S and Leibler RA 1951), which measures the divergence of the true model to the fitted model. Specifically, Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used as the guideline (Burnham and Anderson 1998).

In the Bayesian hierarchical model, computation was done using Markov chain Monte Carlo (MCMC) (Gelfand and Smith 1990) in WinBUGS (Spiegelhalter, Thomas, and Best 2002). Burn-in consisted of 10^5 iterations, and 4×10^5 iterations were used for posterior summaries. Convergence of Markov chains was assessed using the Gelman and Rubin convergence statistic (Gelman and Rubin 1992; Brooks and Gelman 1998). The 95% posterior credible intervals on the transformed scale are available directly from the approximation to the posterior distribution from the MCMC chains. Deviance information criterion (DIC) was used as the guideline to avoid over-fitting the data (Spiegelhalter et al. 2002). The hyper-priors for the precision parameters were assumed to be as follows:

1) $\sigma_{\varepsilon}^{-2} \sim Gamma(1,\ 1)$, which corresponds to a wide 95% confidence interval (CI) of (0.27, 39.50) for the variance parameter σ_{ε}^{2}, allowing large heterogeneity for the prevalence; 2) $\left(\sigma_{\mu_A}^{-2},\sigma_{\mu_B}^{-2},\sigma_{\nu_A}^{-2},\sigma_{\nu_B}^{-2}\right) \sim Gamma(2,\ 2)$, which corresponds to a 95% CI of (0.36, 8.26) for the variance parameters $\left(\sigma_{\mu_A}^{2},\sigma_{\mu_B}^{2},\sigma_{\nu_A}^{2},\sigma_{\nu_B}^{2}\right)$, providing moderate heterogeneity for the latent sensitivities and specificities. Vague priors of N(0, 2^2) were assumed for the fixed parameters $(\eta_0,\alpha_A,\alpha_B,\beta_A,\beta_B)$, which corresponds a 95% CI of log-odds ranging from 0.02 to 50 (Chu et al. 2006). A vague prior of N(0, 2^2) is used for η_1 on the log scale to ensure the constraint that the prevalence of the high risk group is greater than that in the low risk group for any study *i* (Chen, Watson, and Parmigiani 2005). We selected the above non-diffuse priors instead of diffuse priors since the latter can lead to inaccurate posterior estimates (Natarajan and McCulloch 1998).

4. Results of the case study

To identify the best fitting model, we started with the model presented in Chen et al. (2005), which assumed no random effects (referred as Model I). Using the forward-selection procedure presented in Section 3.3, Table 3 presents the goodness of fit statistics using the twice negative log-likelihood, AIC and BIC for the non-linear random effects model and DIC for the Bayesian hierarchical model. At each forward step, we add a random-effects component that provides the largest improvement based on the above model selection criteria.

In the first step, adding any random-effects improved the goodness of fit under all criteria, with the exception of Model IIc using DIC. The largest improvement was achieved by allowing for study-specific prevalence ε_i, referred as Model IIa. For exam-

ple, the DIC decreased by 69.4 points compared to Model I. This revealed an important characteristic of this meta-analysis, that is, the studies varied considerably in their criteria for selecting individuals to be tested. As a result, the study-specific mutation prevalence differed across studies. Based on the Bayesian hierarchical Model IIa, the posterior means of prevalence ranged from 0.125 to 0.860 for the high-risk group, and from 0.016 to 0.098 for the low-risk group.

In the second step, the largest improvement (under all criteria) was seen by adding a random-effects component for mutation analysis sensitivity μ_{iB}, referred as Model IIIc. Although the improvement was modest compared to adding the first random-effects, but it is still notable (e.g., the DIC decreased by 15.3 points compared to Model IIa). This is plausible because each study was conducted in a different laboratory using a variety of mutation analysis techniques. As a result, the mutation analysis sensitivities ranged from 0.424 to 0.871 based on the Bayesian hierarchical Model IIIc.

The last step that introduced any improvement was by adding random effects for microsatellite instability testing sensitivity μ_{iA}, referred as Model IVa. The DIC decreased by 9.6 points compared to Model IIIc. Thus, the final model included the random-effects on: 1) prevalence ε_i; 2) mutation analysis sensitivity μ_{iB}; and 3) microsatellite instability testing sensitivity μ_{iA}. It is worth noting that the model selection proceeded identically under the non-linear mixed effects model and the Bayesian hierarchical model in this case study.

For comparison, the main effects (MSI sensitivity, MSI specificity, MUT sensitivity, MUT specificity, prevalence in the high risk group and prevalence in the low risk group) from Model I, Model IIa, Model IIIc and Model Iva were presented in Table 4 us-

ing the non-linear mixed effects models and the Bayesian hierarchical models. The main effects estimates were highly concordant between the two approaches, except some difference in the estimates of MSI sensitivity. Based on the final model, the posterior median of MSI sensitivity was 0.92 with 95% equal tail credible sets (0.74, 0.99), while the point estimate from the non-linear random effects model was 0.97 with 95% confidence interval (0.92, 1.00).

The Bayesian posterior means with 95% equal tail credible sets of the study-specific effects from the final model are shown in Figure 1. The study-specific MSI sensitivity estimates were quite homogeneous, with study 13 being the only exception (See Figure 1C), which is consistent with the expert belief that MSI is a relatively standard and simple test and the measurement variability associated with it is low. A possible explanation for the exception is that the missense mutations found in those studies may be non-pathogenic, such that the tumors emerged through a pathway different from the MMR mechanism thus did not exhibit the MSI characteristic (Samowitz et al. 2002). On the other hand, the study-specific estimates of mutation prevalence and MUT sensitivity are quite heterogeneous suggesting difference in the selection of study populations and in the laboratory work for MUT test.

Figure 2 presents the posterior kernel smoothed density of MSI sensitivity, MSI specificity, MUT sensitivity and MUT specificity based on the final Bayesian hierarchical model IVa using 4×10^5 Monte Carlo samples, suggesting a very skewed posterior density of MSI sensitivity, which may partially explain the difference in MSI sensitivity estimates between the non-linear mixed effect model and the Bayesian hierarchical model.

5. Simulation studies

Four sets of simulations with 2000 replications each were performed to evaluate the sampling properties of potential misspecification of random effects in the estimation of MSI sensitivity. To reduce the complexity of selecting potential random effects and computational time, we only considered random effects on the disease prevalence and/or test sensitivities $(\varepsilon_i, \mu_{iA}, \mu_{iB})$ and only fitted models with up to two random effects.

Specifically, data were generated from: 1) no random effects; 2) random effects on prevalence (ε_i); 3) random effects on MSI sensitivity (μ_{iA}); and 4) random effects on prevalence and MSI sensitivity (ε_i, μ_{iA}). For each simulation, 20 meta-studies were generated with 7 studies having a high risk group, 7 studies having a low risk group, and 6 studies having both high-risk and low-risk groups. Eighty observations for the high risk group and 250 for the low risk group were generated, which roughly match the average sample sizes per study group in the case study. Each low risk group has a probability of 0.40 to only conduct MUT testing among those with positive MSI test. It is a common scenario in diagnostic testing literature when the reference test (MUT) is expensive or invasive. In the absence of random effects, the prevalence of true mutation were set to be 50% for the high risk group and 10% for the lower risk group, the sensitivity and specificity were both 90% for MSI testing, and 70% and 98% for MUT analysis, respectively. In the presence of random effects, the variances of $(\varepsilon_i, \mu_{iA}, \mu_{iB})$ were set to be 0.5^2 which gives the prevalence a 95% CI of 27-73% for the high risk group and 4-23% for the low risk group, and the MSI sensitivity a 95% CI of 77-96%.

For each generated dataset, we fit seven models using both NLMIXED and BHM: 1) no random effect; 2) one random effect (on ε_i, μ_{iA} or μ_{iB}); and 3) two random effects (on $[\varepsilon_i, \mu_{iA}]$, $[\varepsilon_i, \mu_{iB}]$, or $[\mu_{iA}, \mu_{iB}]$). Table 5 summarizes the Monte Carlo frequency of se-

lecting each candidate model as the "best" model for each set of simulations. Model selection was based on AIC and BIC for the non-linear random effects model using SAS PROC NLMIXED and DIC for the Bayesian hierarchical model using WinBUGs. In summary, DIC has a probability of 0.55-0.70 to identify the true random effects model, while the performance of AIC and BIC is highly variable with a probability of 0.25-0.95. Closer examination of the results reveals that the Bayesian approach (DIC) has a stronger tendency to select additional random effect(s) not included in the true model than does the non-linear random effects approach (overall probability of 0.17 for DIC, 0.06 for AIC and 0.03 for BIC averaging overall four scenarios). Similarly, the overall average probability that the Bayesian approach fails to find true random effects is lower than the non-linear random effects approach (0.17 based on DIC, 0.30 based on AIC and 0.36 based on BIC). For the random effect in the prevalence (ε_i), it is almost always identified if present. Under-fitting are mainly due to the failure to include the random effect in MSI sensitivity (μ_{iA}), which has a narrower distribution than ε_i by simulation design due to the logit transformation. Overall, the probability of selecting completing incorrect random effects (i.e. including fake random effects while failing to include true random effects) is very low under all criteria (0.03 for DIC, 0.03 for AIC, 0.01 for BIC, respectively).

Table 6 records the means, the standard errors, 95% CI lengths and 95% CI coverage probabilities for the MSI sensitivity based on each candidate model using the nonlinear random effects model and the Bayesian hierarchical model, respectively. The coverage probabilities are all close to the nominal value of 0.95 under the true model. When over-fitting occurs, the coverage probabilities are not much affected. When under-fitting or mis-fitting occurs, failure to include random effects in prevalence (ε_i) does not

affect the coverage probabilities for the MSI sensitivity; however failure to include random effects in MSI sensitivity (μ_{iA}) will have a less than nominal coverage. It emphasizes the need to carefully select a random effects model on the estimation of diagnostic accuracy measurements from a meta-analysis without a gold standard.

6. Discussion

In this paper, we demonstrate an improved estimation of the sensitivity and specificity of MSI and traditional mutation analysis by using a non-linear random effects model and a Bayesian hierarchical model from a meta-analysis without a gold standard in the presence of missing data, which has taken the heterogeneity across studies into consideration through study-specific random effects. Simulation studies have demonstrated the importance of carefully selecting appropriate random effects on the estimation of diagnostic accuracy measurements in this scenario. The proposed methods can be used to estimate the accuracy of two imperfect diagnostic tests in other meta-analyses when the prevalence of disease, the sensitivities and/or the specificities of diagnostic tests may be heterogeneous across studies.

In this meta-analysis, the model selection criteria consistently showed that allowing for the appropriate random effects improves the goodness of fit. This also made an impact to the estimates of the parameters of interest, the sensitivity and specificity of MSI and MUT. In particular, the MSI sensitivity estimate has increased noticeably. It was believed that all tumors from Lynch syndrome individuals (i.e. mutation carriers) exhibit MSI, only a small fraction may show a low level of MSI, or MSI-L, which is categorized into missense mutations in this meta-analysis according to conventions. Therefore, the new higher estimate is more biologically plausible (Vasen and Boland 2005). Further-

more, random effects models may be useful in identifying studies that are outliers. For example, the missense mutations found in study 13 might be non-pathogenic and did not exhibit the MSI characteristic such that MSI had a very low sensitivity.

When estimating random effects in the presence of frequent missing data, the convergence may become an issue. For example, we were not able to fit the non-linear mixed effects model with random effects on prevalence, MSI sensitivity, MSI specificity, MUT sensitivity and MUT specificity simultaneously using PROC NLMIXED. Furthermore, about 0.1-0.5% simulations did not converge when using PROC NLMIXED with starting values set to be the true parameters.

In this case study, we assumed independence between MSI test result and mutation analysis result given the true mutation status conditional on the study-specific random effects, which is biologically plausible since large genomic deletions and rearrangements do not differ from the other mutations in their ability to generate microsatellite instable tumors. If the conditional independence assumption is suspicious, methods incorporating dependent errors need to be considered (Torrance-Rynard and Walter 1997; Dendukuri and Joseph 2001). However, given the complexity of meta-analysis (e.g., heterogeneity across studies and missing data due to partial testing), further research is needed on how to incorporate dependent errors.

References

1. Andersen S (1997) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 145 (3):290.
2. Brooks SP and Gelman A (1998) Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7:434-455.

3. Burnham KP, Anderson DR (1998) Model Selection and Inference: A Practical Information-Theoretic Approach. Springer-Verlag: New York.

4. Carlin BP, Louis TA (2000) Bayes and Empirical Bayes Methods for Data Analysis. 2nd ed. Chapman & Hall/CRC.

5. Chen S, Watson P, and Parmigiani G (2005) Accuracy of MSI testing in predicting germline mutations of MSH2 and MLH1: a case study in Bayesian meta-analysis of diagnostic tests without a gold standard. *Biostat* 6 (3):450-464.

6. Chu H et al (2006) Sensitivity analysis of misclassification: a graphical and a Bayesian approach. *Ann Epidemiol* 16:834-841.

7. Davidian M, Giltinan DM (1995) Nonlinear models for repeated measurement data. 1st ed. Chapman & Hall/CRC: Boca Raton.

8. Dendukuri N and Joseph L (2001) Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 57 (1):158-167.

9. Gart JJ and Buck AA (1966) Comparison of A Screening Test and A Reference Test in Epidemiologic Studies .2. A Probabilistic Model for Comparison of Diagnostic Tests. *American Journal of Epidemiology* 83 (3):593-&.

10. Gelfand AE and Smith AFM (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85 (410):398-409.

11. Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian Data Analysis. Chapman & Hall/CRC.

12. Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* 138:182-195.

13. Halloran ME, Preziosi MP, and Chu HT (2003) Estimating vaccine efficacy from secondary attack rates. *Journal of the American Statistical Association* 98 (461):38-46.

14. Hui SL and Walter SD (1980) Estimating the Error Rates of Diagnostic-Tests. *Biometrics* 36 (1):167-171.

15. Johnson WO, Gastwirth JL, and Pearson LM (2001) Screening without a "gold standard": The Hui-Walter paradigm revisited. *American Journal of Epidemiology* 153 (9):921-924.

16. Joseph L, Gyorkos TW, and Coupal L (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am.J.Epidemiol.* 141 (3):263-272.

17. Kullback S and Leibler RA (1951) On information and sufficiency. *Annals of Mathematical Statistics* 22:79-86.

18. Little RJA, Rubin DB (2002) Statistical Analysis With Missing Data. 2nd ed. John Wiley & Sons.

19. Lynch HT and de la Chapelle A (1999) Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* 36 (11):801-818.

20. Molenberghs G, Verbeke G (2005) Models for Discrete Longitudinal Data. Springer.

21. Natarajan R and McCulloch CE (1998) Gibbs sampling with diffuse proper priors: a valid approach to data-driven inference? *Journal of Computational and Graphical Statistics* 7 (3):267-277.

22. Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press: Oxford.

23. Pinheiro JC and Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 4 (1):12-35.

24. Samowitz WS et al (2002) Missense Mismatch Repair Gene Alterations, Microsatellite Instability, and Hereditary Nonpolyposis Colorectal Cancer. *J Clin Oncol* 20 (14):3178-3179.

25. Spiegelhalter DJ et al (2002) Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, Series B* 63 (4):583-639.

26. Spiegelhalter, D. J., Thomas, A., and Best, N. G. WinBUGS user manual, version 1.4. 2002.
Ref Type: Unpublished Work

27. Torrance-Rynard VL and Walter SD (1997) Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* 16 (19):2157-2175.

28. Umar A et al (2004) Revised Bethesda Guidelines for Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndrome) and Microsatellite Instability. *J Natl Cancer Inst* 96 (4):261-268.

29. Vasen HFA and Boland CR (2005) Progress in Genetic Testing, Classification, and Identification of Lynch Syndrome. *JAMA* 293 (16):2028-2030.

30. Vonesh EF, Chinchilli VM (1997) Linear and Nonlinear Models for the Analysis of Repeated Measurements. Marcel Dekker: New York.

31. Zhou XH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. John Wiley & Sons: New York.

Table 1. A list of the studies included in the Meta-analysis.

Study	ID	High Risk	n_{i11}	n_{i10}	n_{i01}	n_{i00}	n_{i1m}	n_{i0m}	n_{im1}	n_{im0}
Bapat et al. (1999)	1	Y	16	1	2	20	0	0	0	0
Calistri et al. (2000)	2	Y	8	8	0	9	0	0	0	0
Cederquist *et al.* (2001)	3	Y	8	15	0	0	12	43	0	0
Debniak et al. (2000)	4	Y	5	4	1	15	0	0	0	0
Debniak et al. (2000)	4	N	0	5	0	38	0	0	0	0
Dietmaier *et al.* (1997)	5	N	0	0	0	0	18	130	0	0
Dieumegard et al. (2000)	6	Y	7	7	0	2	0	0	0	0
Dieumegard et al. (2000)	6	N	0	0	0	7	0	0	0	0
Lamberti et al. (1999)	7	Y	13	22	2	10	0	0	0	0
Liu et al. (2000)	8	Y	16	6	1	36	0	0	0	0
Loukola et al. (1999)	9	N	10	53	0	0	0	446	0	0
Percesepe et al. (2001)	10	N	0	0	0	0	28	308	0	0
de Leon *et al*. (2004)	11	Y	0	0	0	0	0	0	89	75
Salahshor *et al*. (1999)	12	N	0	0	0	0	22	159	0	0
Scartozzi *et al*. (2002)	13	Y	0	1	4	22	0	0	0	0
Salovaara *et al*. (2000)	14	N	18	48	0	0	0	469	0	0
Terdiman *et al*. (2001)	15	Y	21	11	0	0	0	63	0	0
Wahlberg *et al*. (2002)	16	Y	14	14	0	20	0	0	0	0
Wang *et al*. (2003)	17	Y	92	88	0	0	0	188	0	0

Note: $(n_{i11}, n_{i10}, n_{i01}, n_{i00}, n_{i1m}, n_{i0m}, n_{im1}, n_{im0})$ correspond to the number of subjects with MSI = 1 & MUT = 1, MSI = 1 & MUT = 0, MSI = 0 & MUT = 1, MSI = 0 & MUT = 0, MSI = 1 & MUT = missing, MSI = 0 & MUT = missing, MSI = missing & MUT = 1, and MSI = missing & MUT = 0, respectively. MSI = microsatellite instability testing, MUT = mutation analysis testing.

Table 2. Typical data displays for study i ($i = 1, \ldots, I$) with missing data.

MSI	MUT		
	Positive (+)	Negative (–)	Missing
Positive (+)	n_{i11} $(1-P_{iA}-P_{iB})P_{i11}$	n_{i10} $(1-P_{iA}-P_{iB})P_{i10}$	n_{i1m} $P_{iA}(P_{i11}+P_{i10})$
Negative (–)	n_{i01} $(1-P_{iA}-P_{iB})P_{i01}$	n_{i00} $(1-P_{iA}-P_{iB})P_{i00}$	n_{i0m} $P_{iA}(P_{i01}+P_{i00})$
Missing	n_{im1} $P_{iB}(P_{i11}+P_{i01})$	n_{im0} $P_{iB}(P_{i10}+P_{i00})$	—

Note: Probabilities corresponding to a given cell are shown in the second line.
MSI = microsatellite instability testing, MUT = mutation analysis testing.

Table 3. Selection of random effects using a forward selection procedure

Random Effects Models	NLMM Using NLMIXED			BHM Using WinBUGS	
	-2logL	AIC	BIC	DIC	p_D
I	**91.3**	**103.3**	**108.3**	**104.2**	**5.6**
IIa (ε_i)	**44.5**	**58.5**	**64.4**	**34.8**	**17.7**
IIb (μ_{iA})	67.9	81.9	87.7	67.0	14.5
IIc (ν_{iA})	81.7	95.7	101.5	119.9	10.8
IId (μ_{iB})	64.0	78.0	83.8	65.0	14.6
IIe (ν_{iB})	66.0	80.0	85.8	68.4	8.5
IIIa (ε_i, μ_{iA})	39.8	55.8	62.4	23.1	15.9
IIIb (ε_i, ν_{iA})	44.2	60.2	66.9	24.3	17.3
IIIc (ε_i, μ_{iB})	**36.8**	**52.8**	**59.4**	**19.5**	**24.8**
IIId (ε_i, ν_{iB})	42.6	58.6	65.3	27.3	17.8
IVa (ε_i, μ_{iB}, μ_{iA})	**30.6**	**48.6**	**56.1**	**9.9**	**24.0**
IVb (ε_i, μ_{iB}, ν_{iB})	34.9	52.9	60.4	16.1	21.2
IVc (ε_i, μ_{iB}, ν_{iA})	36.5	54.5	62.0	14.8	26.2
IVd (ε_i, μ_{iB}, μ_{iA}, ρ_μ)	30.9	50.9	59.2	14.2	24.2
IVe (ε_i, μ_{iB}, ν_{iB}, ρ_B)	34.6	54.6	63.0	22.7	25.9
IVf (ε_i, μ_{iB}, ν_{iA}, $\rho_{\nu_A\mu_B}$)	36.7	56.7	65.0	43.9	27.8
Va (ε_i, μ_{iB}, μ_{iA}, ν_{iA})	31.0	51.0	59.3	12.1	24.9
Vb (ε_i, μ_{iB}, μ_{iA}, ν_{iB})	30.9	50.9	59.2	8.5	23.0

*Test A and B correspond to the microsatellite instability (MSI) and mutation analysis (MUT) testing, respectively. NLMM = non-linear mixed effects model; BHM = Bayesian hierarchical model; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; DIC = Deviance Information Criterion; p_D = the effective number of parameters. For the Bayesian analysis, priors for precision parameters of random effects are specified as $\sigma_\varepsilon^{-2} \sim \text{Gamma}(1, 1)$ and $\left(\sigma_{\mu_A}^{-2}, \sigma_{\mu_B}^{-2}, \sigma_{\nu_A}^{-2}, \sigma_{\nu_B}^{-2}\right) \sim Gamma(2, 2)$. The random effects $(\varepsilon_i, \mu_{iA}, \nu_{iA}, \mu_{iB}, \nu_{iB})$ correspond to study-specific prevalence, MSI sensitivity, MSI specificity, MUT sensitivity and MUT specificity, respectively. Thirty-three hundred points have been deducted from -2logL, AIC, BIC and DIC for presentation.

Table 4. Summary of parameter estimates using the non-linear random effects models and the Bayesian hierarchical models: the triple notation of ${}_{L}P_{U}$ denotes the point estimate P with 95% confidence limits (L, U) for the non-linear random effects models, or the posterior median P with 95% equal tailed credible limits (L, U) using Bayesian hierarchical models. The numbers have been multiplied by 1000 for presentation.

	Non-linear Random Effects Models* Using NLMIXED				Bayesian Hierarchical Models Using WinBUGS			
Random Effects Models	I	IIa	IIIc	IVa	I	IIa	IIIc	IVa
	None	ε_i	ε_i , μ_{iB}	ε_i , μ_{iB} , μ_{iA}	None	ε_i	ε_i , μ_{iB}	ε_i , μ_{iB} , μ_{iA}
MSI Specificity	${}_{902}920_{937}$	${}_{893}912_{932}$	${}_{889}909_{929}$	${}_{894}914_{934}$	${}_{898}917_{936}$	${}_{898}916_{934}$	${}_{893}912_{938}$	${}_{895}914_{939}$
MSI Sensitivity	${}_{735}819_{904}$	${}_{892}978_{1000}$	${}_{880}982_{1000}$	${}_{922}968_{1000}$	${}_{745}842_{951}$	${}_{843}934_{985}$	${}_{872}957_{996}$	${}_{740}922_{990}$
MUT Specificity	${}_{1000}1000_{1000}$	${}_{906}953_{999}$	${}_{898}952_{1000}$	${}_{947}986_{1000}$	${}_{917}980_{998}$	${}_{926}968_{996}$	${}_{916}958_{990}$	${}_{935}981_{998}$
MUT Sensitivity	${}_{564}622_{679}$	${}_{536}594_{653}$	${}_{508}656_{805}$	${}_{495}641_{786}$	${}_{565}621_{678}$	${}_{537}590_{645}$	${}_{531}630_{726}$	${}_{488}645_{801}$
High-risk Prevalence	${}_{491}555_{618}$	${}_{318}495_{673}$	${}_{302}465_{629}$	${}_{442}532_{622}$	${}_{460}536_{610}$	${}_{354}520_{694}$	${}_{328}476_{641}$	${}_{380}532_{685}$
Low-risk Prevalence	${}_{28}47_{65}$	${}_{0}18_{43}$	${}_{0}17_{39}$	${}_{8}28_{47}$	${}_{24}42_{66}$	${}_{9}29_{75}$	${}_{9}27_{72}$	${}_{11}33_{82}$
σ_{ε} (prevalence)	—	${}_{445}1034_{1624}$	${}_{354}904_{1454}$	${}_{311}601_{890}$	—	${}_{672}1060_{1790}$	${}_{605}959_{1626}$	${}_{497}798_{1384}$
σ_{μ_A} (sensitivity)	—	—	—	${}_{880}2529_{4177}$	—	—	—	${}_{737}1649_{3766}$
σ_{ν_A} (specificity)	—	—	—	—	—	—	—	—
σ_{μ_B} (sensitivity)	—	—	${}_{111}742_{1373}$	${}_{137}756_{1375}$	—	—	${}_{601}944_{1650}$	${}_{597}932_{1610}$
σ_{ν_B} (specificity)	—	—	—	—	—	—	—	—

*95% confidence intervals based on normal approximation.

Table 5. The empirical probability of selecting a candidate random effects model as the final model using AIC, BIC or DIC* based on simulation studies with 2000 replicates. The bolded cells represent the probability of identifying correctly specified random effects model. The numbers have been multiplied by 1000 for presentation.

True Random Effects Model		Selected Random Effects Model						
		None	ε_i	μ_{iA}	μ_{iB}	ε_i, μ_{iA}	ε_i, μ_{iB}	μ_{iA}, μ_{iB}
None	AIC	**885**	30	41	42	1	1	1
	BIC	**940**	16	21	24	0	0	0
	DIC	**707**	21	188	58	7	3	18
ε_i	AIC	1	**914**	0	1	34	50	1
	BIC	2	**961**	0	1	16	21	0
	DIC	0	**701**	1	1	200	97	1
μ_{iA}	AIC	602	24	**321**	31	9	1	12
	BIC	711	14	**247**	19	4	0	5
	DIC	335	13	**554**	29	20	2	49
ε_i, μ_{iA}	AIC	0	632	0	0	**324**	44	1
	BIC	1	731	1	1	**248**	20	0
	DIC	0	330	1	0	**605**	64	1

*AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; DIC = Deviance Information Criterion.

Table 6. The impact of using different random effects models on MSI sensitivity (true value = 0.90) based on simulation studies with 2000 replicates. The numbers have been multiplied by 1000 for presentation. The bolded cells represent the estimates from a model with correctly specified random effects.

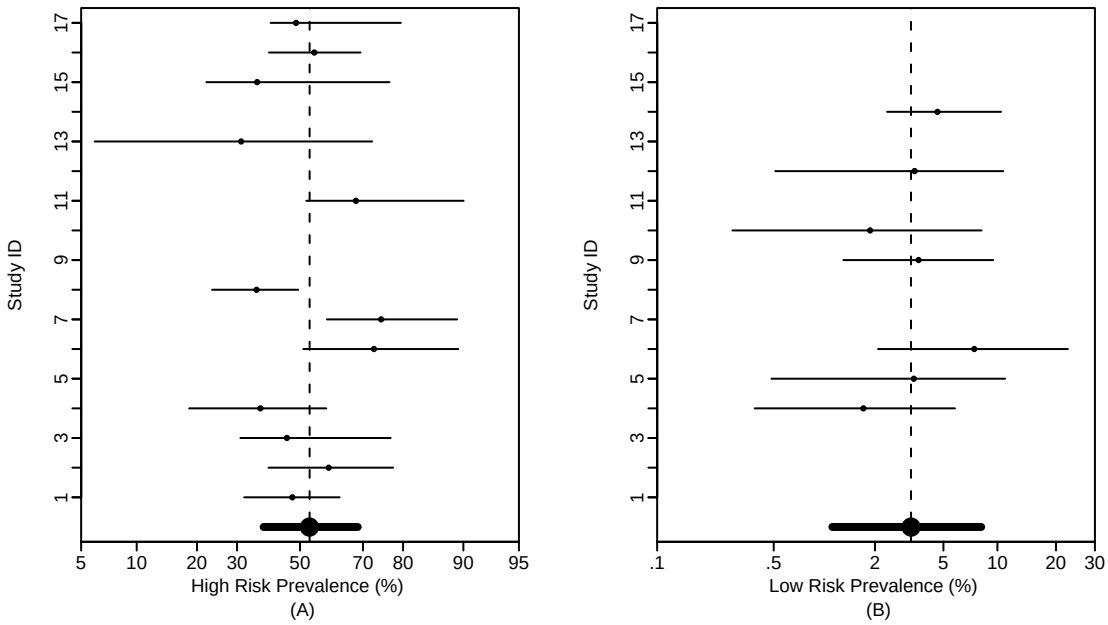
True Models		Random Effects Models Using NLMIXED							Bayesian Hierarchical Models Using WinBUGS						
		None	ε_i	μ_{iA}	μ_{iB}	ε_i, μ_{iA}	ε_i, μ_{iB}	μ_{iA}, μ_{iB}	None	ε_i	μ_{iA}	μ_{iB}	ε_i, μ_{iA}	ε_i, μ_{iB}	μ_{iA}, μ_{iB}
None	Mean	**902**	900	903	900	901	900	902	**897**	900	919	902	918	902	922
	Std Err	**21**	21	23	21	22	021	22	**20**	21	27	21	26	21	26
	95% CI length*	**83**	89	98	89	96	89	97	**79**	84	105	83	101	81	101
	95% CICP*	**961**	968	975	964	977	969	976	**938**	949	944	951	942	946	920
ε_i	Mean	902	**900**	897	896	902	900	890	897	**900**	902	897	916	902	896
	Std Err	21	**20**	28	22	22	20	29	21	**21**	32	22	27	20	33
	95% CI length*	85	**86**	123	92	96	87	129	80	**81**	126	86	103	79	130
	95% CICP*	961	**966**	972	956	977	968	956	941	**958**	981	946	950	945	977
μ_{iA}	Mean	893	891	**900**	891	898	891	899	888	892	**911**	894	911	895	915
	Std Err	21	21	**26**	21	25	21	26	21	22	**29**	22	28	21	28
	95% CI length*	85	91	**111**	91	111	91	111	81	86	**112**	85	108	83	108
	95% CICP*	864	879	**949**	873	942	876	950	829	879	**959**	890	951	875	948
ε_i, μ_{iA}	Mean	893	892	890	887	**899**	892	882	889	893	894	888	**909**	895	887
	Std Err	21	21	32	22	**25**	21	33	21	21	34	22	**28**	21	35
	95% CI length*	85	86	132	92	**106**	87	139	82	84	134	88	**110**	81	138
	95% CICP*	860	885	944	852	**952**	886	898	831	873	977	855	**957**	866	963

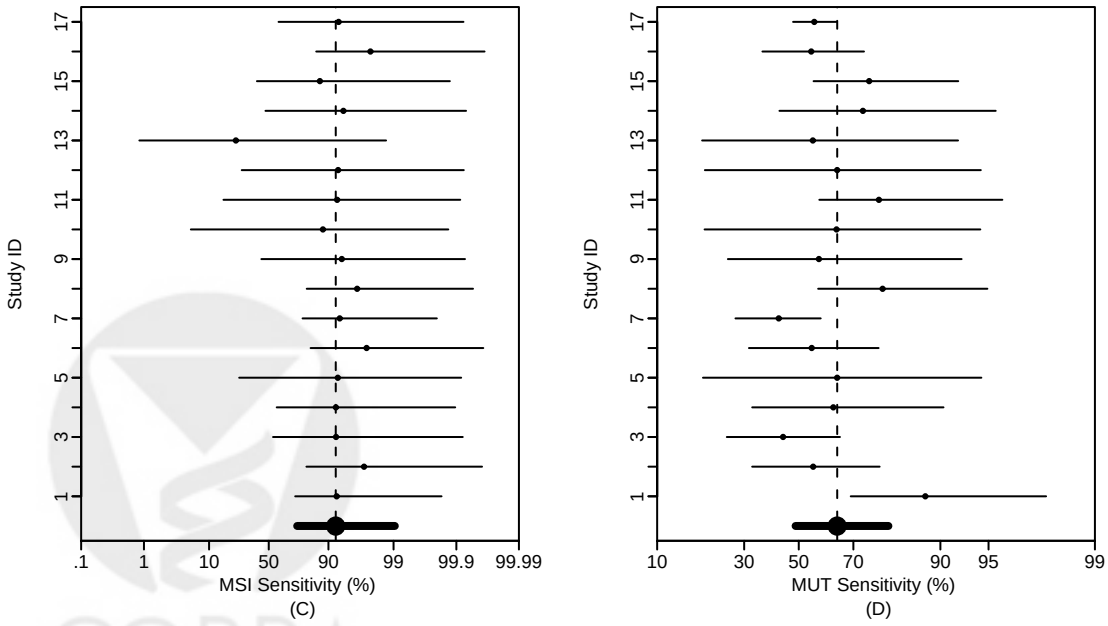
*95% CICP = 95% confidence interval coverage probability, 95% CICP and 95% CI length are based on logit-normal assumption for the random effects models using NLMIXED and equal tail credible intervals for the Bayesian hierarchical models using WinBUGS.

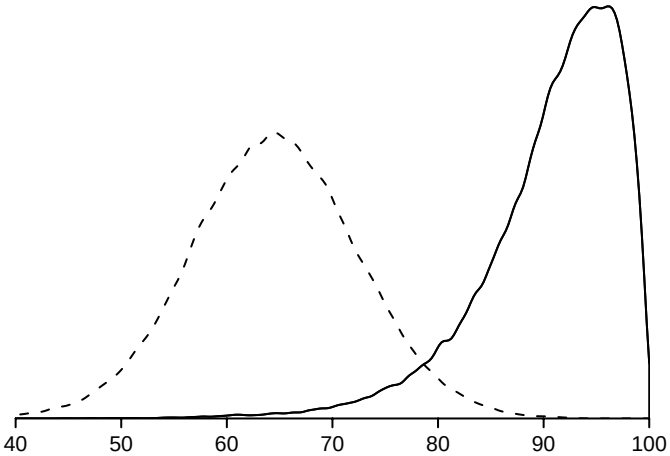
Captions:

Figure 1. Study-specific posterior means with 95% equal tail credible sets of the prevalence of high (panel A) and low (panel B) risk groups, MSI (panel C) and MUT (panel D) sensitivities based on the Bayesian hierarchical model IVa. Bold lines are population averaged posterior estimates (see Section 3.2 for computation of the population averaged values).

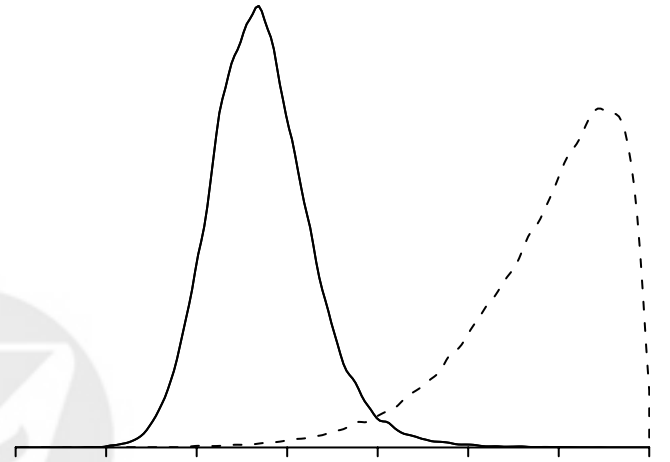
Figure 2. Posterior distributions of MSI and MUT sensitivities (panel A), MSI and MUT specificities (panel B). It is based on the kernel smoothed density estimation of 4×10^5 Monte Carlo samples. Solid and dashed lines denote MSI and MUT, respectively.







A: Sensitivities of MSI and MUT (%)



B: Specificities of MSI and MUT (%)