Web-based supplementary materials for "Estimating diagnostic accuracy of raters without a gold standard by exploiting a group of experts"

by BO ZHANG[1,2], ZHEN CHEN[1], and PAUL S. ALBERT[1]

[1]Biostatistics and Bioinformatics Branch, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD 20892, U.S.A.

[2]Biostatistics Core, School of Biological and Population Health Sciences, College of Public Health and Human Sciences, Oregon State University, OR 97331, U.S.A.

Web Appendix A. Summary of the Data from the Physician Reliability Study

As a summary of the data from the PRS, we present the scatterplot of IE's ratings (in sum) versus R-OB/GYNs ratings (in sum) in Figure S.1.

Web Appendix B. Maximum likelihood estimation: a Monte-Carlo EM algorithm

We consider a Monte-Carlo EM (MCEM) algorithm (McCulloch, 1997; Booth and Hobert, 1999) to obtain the maximum likelihood estimation in Equation (6) in the article as follows. The maximum likelihood estimation of Equation (3) in the article can be similarly derived. We treat the latent true disease status $D = (D_1, \cdots, D_I)'$, random effects

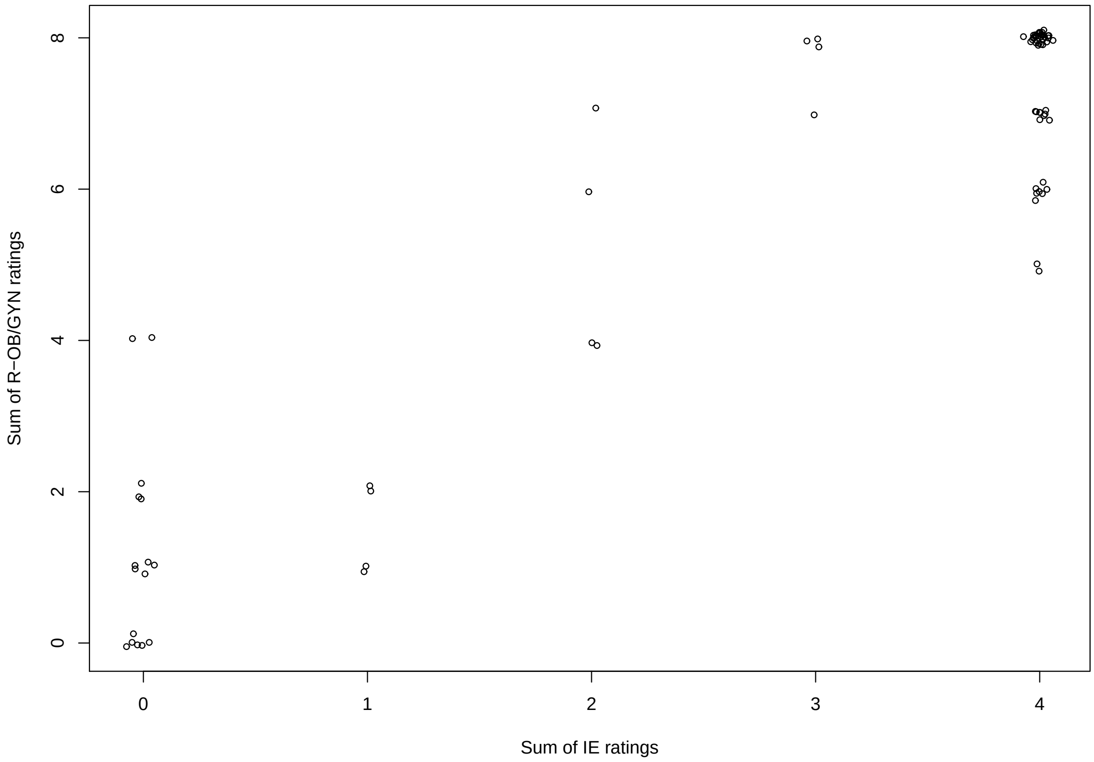

Figure S.1: Subject-wise scatterplot of the sum of the eight OB/GYN ratings versus the sum of the four IE ratings. The points have been jittered to avoid overlapping.

$b = (b_1, \cdots, b_I)'$ and $c = (c_1, \cdots, c_J)'$ as missing data in the EM algorithm. We denote $X^* = (Y_1', \cdots, Y_I', T_1, \cdots, T_I)'$ as the observed data, $Z^* = (D', b', c')'$ as the missing data, and $Y^* = (X^*, Z^*)'$ as the complete data. At the $(r+1)$th iteration of EM algorithm, the E-step involves calculation of the Q-function $Q(\theta|\theta^{(r)}) = E[\log f(Y^*|\theta)|x^*, \theta^{(r)}] = \int f(z^*|x^*, \theta^{(r)}) \log f(y^*|\theta) \mathrm{d}z^*$, where $\theta^{(r)}$ denotes the parameter value from the rth iteration, $f(z^*|x^*, \theta^{(r)})$ is the conditional distribution of missing data given the observed data and $\theta^{(r)}$, and $f(y^*|\theta)$ is the full likelihood, whose logarithm is

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\Bigg[y_{ij}\log\Phi\bigg(\beta_{d_i}+\sigma_{d_i}b_i+\tau_{d_i}c_j\bigg)+(1-y_{ij})\log\bigg\{1-\Phi\bigg(\beta_{d_i}+\sigma_{d_i}b_i+ \tau_{d_i}c_j\bigg)\bigg\}\Bigg]+\sum_{i=1}^{I}\log(S^T_{t_i|d_i})+\sum_{i=1}^{I}\log(\pi_{d_i})+\sum_{i=1}^{I}\log\Big\{g_1(b_i)\Big\}+\sum_{j=1}^{J}\log\Big\{g_2(c_j)\Big\}. \tag{1}$$

The M-step involves maximizing $Q(\theta|\theta^{(r)})$ with respect to θ to yield the new update $\theta^{(r+1)}$. The process is iterated from a starting value $\theta^{(0)}$ to convergence. Under regularity conditions, the value at convergence maximizes the likelihood function. The dimensionality of integration and summations in the likelihood function and in $f(y^*|\theta)$ increase with the number of tests and study subjects so that the integration in the E-step is intractable with numerical integration techniques. Thus, we conduct estimation using MCEM algorithm. Monte-Carlo approximation is formed in the MCEM algorithm to compute the required expectation in $Q(\theta|\theta^{(r)})$. Booth and Hobert (1999) proposed using a rejection or an importance sampling scheme. Their method produces independent and identically distributed samples that may be used to assess Monte-Carlo error at each iteration and hence suggests a rule for changing the sample size to enhance speed. However, their method may break down when the intractable integrals in the likelihood function are of high dimension. An alternative approach is to use the Metropolis-Hastings algorithm, as in McCulloch (1997), to sample from the conditional distribution using the density of unobserved variables as the proposal distribution. In this article, we use the Metropolis-Hastings algorithm, but with an adaptive proposal distribution. To generate N values $z^{*(n)}$, $n = 1, 2, \cdots, N$, from the conditional distribution $f(z^*|x^*, \theta^{(r)})$ using the Metropolis-Hasting algorithm, we start from the starting values $z^{*(0)} = (d^{(0)}, b^{(0)}, c^{(0)})'$ with $f(z^{*(0)}) = f(d^{(0)}, b^{(0)}, c^{(0)}) > 0$. At the nth Metropolis-Hasting step, we sample a candidate $(\tilde{d}, \tilde{b}, \tilde{c})$ from the proposal distribution $\tilde{f}^{(r)}(d, b, c)$. With probability $\min(1, \omega)$, $(d^{(n+1)}, b^{(n+1)}, c^{(n+1)}) = (\tilde{d}, \tilde{b}, \tilde{c})$, where $\omega = f(\tilde{d}, \tilde{b}, \tilde{c}|x^*, \theta^{(r)})\tilde{f}^{(r)}(d^{(n)}, b^{(n)}, c^{(n)})/f(d^{(n)}, b^{(n)}, c^{(n)}|x^*, \theta^{(r)})\tilde{f}^{(r)}(\tilde{d}, \tilde{b}, \tilde{c})$; With probability $1 - \min(1, \omega)$, $(d^{(n+1)}, b^{(n+1)}, c^{(n+1)}) = (d^{(n)}, b^{(n)}, c^{(n)})$. The proposal distribution at the rth iteration of the EM algorithm is constructed to be $\tilde{f}^{(r)}(d, b, c) = \prod_{i=1}^{I} \tilde{f}_D^{(r)}(d_i)\tilde{f}_b^{(r)}(b_i) \prod_{j=1}^{J} \tilde{f}_c^{(r)}(c_j)$, where $\tilde{f}_D^{(r)}(x)$ is the empirical distribution of the d_i's sampled from the $(r-1)$th MCEM iteration and $\tilde{f}_b^{(r)}(x)$ and $\tilde{f}_c^{(r)}(x)$ are two Student's t distribution centered, respectively, at the means of the b_i's and c_j's sampled from the $(r-1)$th MCEM iteration (Gelman, 1995). The proposal distribution here dynamically incorporates the obtained information of the conditional distribution $f(z^*|x^*, \theta^{(r)})$ from the previous MCEM step, and thus greatly improves the performance of the Metropolis-Hasting Markov chain. The starting values $z^{*(0)}$ of the Metropolis-Hastings algorithm are chosen to be the M-step estimates of the previous step of the MCEM algorithm. In the analysis of the PRS study and numerical simulation studies presented in the next two sections, we run 100 steps in the EM algorithm and 10^5 Monte-Carlo iterations in each MCEM step.

The likelihood of our proposed model is intractable due to high-dimensional integration and summation. As a consequence, information criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) is difficult to use to select among candidate latent class models. We choose to use the model selection criteria $\mathrm{IC}_{H,Q}$ proposed by Ibrahim, Zhu, and Tang (2008):

$$\mathrm{IC}_{H,Q} = -2Q(\hat{\theta}|\hat{\theta}) + 2H(\hat{\theta}|\hat{\theta}) + c_{I,n_{par}} \cdot n_{par}, \tag{2}$$

where n_{par} is the number of independent parameters in the candidate model, $c_{I,n_{par}}$ is the penalization term depending on I and n_{par} (for instance, $c_{I,n_{par}} = \log(I)$), $Q(\cdot|\cdot)$ is the Q-function in the EM algorithm, $H(\cdot|\cdot)$ is the H-function in the EM algorithm, and $\hat{\theta}$ is the estimate of parameter θ. In (2), $Q(\hat{\theta}|\hat{\theta}) = \int \log f(y^*|\hat{\theta}) f(z^*|x^*, \hat{\theta}) dz^*$ is a direct byproduct of the MCEM algorithm, since it can be obtained by running one more step after $\hat{\theta}$ is obtained in the MCEM and plugging $\hat{\theta}$ into the Monte-Carlo estimate of $Q(\theta|\hat{\theta})$ from this extra step. The H-function $H(\hat{\theta}|\hat{\theta}) = E[\log f(z^*|x^*, \hat{\theta})] = \int \log f(z^*|x^*, \hat{\theta}) f(z^*|x^*, \hat{\theta}) dz^*$, however, needs to be further estimated because density $f(z^*|x^*, \hat{\theta})$ does not have a closed form and thus can not be estimated by Monte-Carlo approximation. Ibrahim, Zhu, and Tang (2008) proposed a semi-nonparametric truncation estimator for density $f(z^*|x^*, \hat{\theta})$ based upon Hermite series expansion: $\hat{f}_k(z^*|x^*, \hat{\theta}) = P^2(t, \boldsymbol{\psi}, k)\phi(z^*; \hat{\mu}(\hat{\theta}), \hat{\Sigma}(\hat{\theta}))$, where $t = R^{-1}(\hat{\theta})(z^* - \hat{\mu}(\hat{\theta}))$, $\hat{\Sigma}(\hat{\theta}) =$

$R(\hat{\theta})R^T(\hat{\theta})$, $\phi(z^*; \hat{\mu}(\hat{\theta}), \hat{\Sigma}(\hat{\theta}))$ is a multivariate normal density with mean $\hat{\mu}(\hat{\theta})$ and covariance matrix $\hat{\Sigma}(\hat{\theta})$, $\hat{\mu}(\hat{\theta})$ and $\hat{\Sigma}(\hat{\theta})$ are the conditional mean and covariance matrix of z^* given x^* and $\hat{\theta}$. Here, $P(t, \psi, k)$ is a multivariate polynomial of order k, and ψ are the coefficient vector of $P(t, \psi, k)$. The coefficient vector ψ is estimated through a set of random sample from $f(z^*|x^*, \hat{\theta})$ via quasi maximum likelihood.

The class of model selection criteria (2) is denoted as $\text{IC}_{H(k),Q}$ if semi-nonparametric estimator $\hat{f}_k(z^*|x^*, \hat{\theta})$ is used to estimate the H-function $H(\hat{\theta}|\hat{\theta})$ in (2). The analytic approximation to the integrand of the H-function and its computation may be cumbersome for large k. But Gallant and Nychka (1987) and Fenton and Gallant (1996) showed that $\hat{f}_k(z^*|x^*, \hat{\theta})$ approximates $f(z^*|x^*, \hat{\theta})$ well for even small k. Ibrahim, Zhu, and Tang (2008) also showed that $\text{IC}_{H(k),Q}$ preforms well with small k. Both the simulation studies and examples in Ibrahim, Zhu, and Tang (2008) demonstrated no much difference between $\text{IC}_{H(0),Q}$ and other $\text{IC}_{H(k),Q}$'s with $k \geq 1$. As a result, we choose to use $\text{IC}_{H(0),Q}$ throughout this paper given its balance between computational complexity and selection accuracy.

Web Appendix C. Simulation results for sensitivity and specificity estimates of R-OB/GYNs from the latent class model that incorporating both R-OB/GYNs and IEs.

Simulation results for sensitivity and specificity estimates of R-OB/GYNs from model (9) in the manuscript are presented in Table S.1.

Table S.1: Simulation results for sensitivity and specificity estimates of R-OB/GYNs from model (9). The averages of estimates (standard errors) and the percentage of selecting true model by $\text{IC}_{H(0),Q}$ are presented. The true sensitivity, specificity and disease prevalence are $S_e = 0.88$, $S_p = 0.87$, and $\pi_1 = 0.7$, respectively.

Number of tests	Working random effects distribution	$\hat{S}_e(se)$	$\hat{S}_p(se)$	$\hat{\pi}_1(se)$	Rate of selecting true model
5	Normal	0.82(0.052)	0.81(0.043)	0.76(0.053)	59%
	MixN	0.88(0.052)	0.86(0.057)	0.69(0.058)	
10	Normal	0.81(0.061)	0.81(0.055)	0.78(0.044)	53%
	MixN	0.88(0.066)	0.87(0.068)	0.70(0.050)	

Web Appendix D. Discussion on the of IE data

The use of IE data is one of the key aspects in the application of the proposed methodology. The following discussion addresses the three issues on the use of IE data.

First, with regard to the number of IEs, it is important to first understand why the inference are not sensitive to parameters of the polychotomous logit model. In part, this is due to the fact that $P(D_i = 1|T_i = 4) \approx 1$ and $P(D_i = 0|T_i = 0) \approx 1$ for all parameters considered. We think this is sensible since it would seem very unlikely that you would have a positive (negative) gold standard when all the IEs were negative (positive). You would need a large enough group of IEs to have this confidence. This number depends on the particular application. However, for a general rule, we recommend a minimum of four expert ratings.

Second, we conducted simulation studies to investigate the performance of the proposed method when IEs only examine a subset of the patients. We repeated the simulation study in Table 1(B) in the manuscript with a subset of the patients examined by the four IEs. Table S.3 in this document shows the simulation results for the scenarios when the IEs examine 80%, 50% and 20% of the patients, respectively. From Table S.3, we can see that the estimates of the sensitivity and specificity of the R-OB/GYNs have no or very little bias when the IEs examine 80% and 50% of the patients. When the proportion of examined patients decreased to 20%, the estimates have more substantical bias.

Table S.2: Simulation results for sensitivity and specificity under the estimated imperfect reference standard when the IEs examine a subset of the patients. The random effects of the true models follow mixture normal (MixN) distribution. The averages of estimates (standard errors) and the percentage of selecting true model by $\mathrm{IC}_{H(0),Q}$ are presented. The true sensitivity, specificity and disease prevalence are $S_e = 0.88$, $S_p = 0.87$, and $\pi_1 = 0.7$, respectively.

Number of tests	Working random effects distribution	$\hat{S}_e(se)$	$\hat{S}_p(se)$	$\hat{\pi}_1(se)$	Rate of selecting true model
IEs rate 80% of the patients					
5	Normal	0.86(0.059)	0.86(0.049)	0.72(0.051)	88%
	MixN	0.88(0.054)	0.87(0.052)	0.69(0.050)	
10	Normal	0.87(0.061)	0.86(0.058)	0.71(0.054)	87%
	MixN	0.88(0.060)	0.87(0.057)	0.70(0.048)	
IEs rate 50% of the patients					
5	Normal	0.87(0.050)	0.86(0.051)	0.71(0.042)	85%
	MixN	0.88(0.055)	0.87(0.057)	0.70(0.051)	
10	Normal	0.87(0.054)	0.85(0.062)	0.71(0.044)	89%
	MixN	0.88(0.061)	0.87(0.054)	0.70(0.053)	
IEs rate 20% of the patients					
5	Normal	0.82(0.063)	0.81(0.061)	0.76(0.046)	62%
	MixN	0.88(0.053)	0.87(0.059)	0.70(0.050)	
10	Normal	0.83(0.068)	0.84(0.067)	0.75(0.063)	59%
	MixN	0.88(0.052)	0.87(0.056)	0.70(0.052)	

Third, we conducted simulation studies to investigate the performance of the proposed method when the patients are not examined by all IEs. We repeated the simulation study in Table 1(B) and Table 1(C) in the manuscript when 20% of the patients miss one of the four IE ratings. We imputed the missing ratings by assuming missing completely at random. Table S.3 in this document shows the simulation results, which indicate that the estimates of the sensitivity and specificity of the R-OB/GYNs are robust. Therefore, we suggest that, when the patients are not examined by all IEs, the proposed method can still function very well with the appropriate imputation for the missingness.

Table S.3: Simulation results for sensitivity and specificity when 20% IE's ratings are missing, under the scenarios (B) with a correctly specified imperfect reference standard ($\gamma_0 = -4.5$, $\gamma_1 = 0.1$, $\gamma_2 = 0.2$ in equations (7) and (8) in the manuscript and (C) with an incorrectly specified imperfect reference standard ($\gamma_0 = -4.5$, $\gamma_1 = 0.1$, $\gamma_2 = 0.1$ equations (7) and (8) in the manuscript. The random effects of the true models follow mixture normal (MixN) distribution. The averages of estimates (standard errors) and the percentage of selecting true model by $\text{IC}_{H(0),Q}$ are presented. The true sensitivity, specificity and disease prevalence are $S_e = 0.88$, $S_p = 0.87$, and $\pi_1 = 0.7$, respectively.

Number of tests	Working random effects distribution	$\hat{S}_e(se)$	$\hat{S}_p(se)$	$\hat{\pi}_1(se)$	Rate of selecting true model
(B)					
5	Normal	0.87(0.057)	0.87(0.067)	0.70(0.053)	94%
	MixN	0.88(0.052)	0.87(0.065)	0.70(0.056)	
10	Normal	0.88(0.064)	0.88(0.051)	0.70(0.057)	95%
	MixN	0.88(0.052)	0.87(0.050)	0.70(0.049)	
(C)					
5	Normal	0.88(0.063)	0.86(0.058)	0.71(0.045)	91%
	MixN	0.88(0.056)	0.87(0.059)	0.70(0.047)	
10	Normal	0.87(0.055)	0.87(0.056)	0.69(0.053)	91%
	MixN	0.88(0.057)	0.87(0.050)	0.70(0.041)	

Web Appendix E. On IE exchangeability

In the Physician Reliability Study (PRS), there are four international experts (IEs). The ratings from these four IEs were taken as the imperfect reference standard in diagnosing endometriosis. Let us denote $e_l = 0$ or 1, $l = 1, 2, 3, 4$, the rating from the lth IE. Under exchangeability of the raters, $P(\tilde{T}_i^{(1)} = e_1, \tilde{T}_i^{(2)} = e_2, \tilde{T}_i^{(3)} = e_3, \tilde{T}_i^{(4)} = e_4 | D_i = d_i) = P(T_i = \sum_{l=1}^{4} e_l | D_i = d_i)$ for any combination of e_1, e_2, e_3, and e_4, where $d_i = 0$ or 1. Thus, when we assume exchangeability between raters, it is reasonable to use the sum of the IE ratings that is characterized by polychotomous logit model in Page 8 in the manuscript.

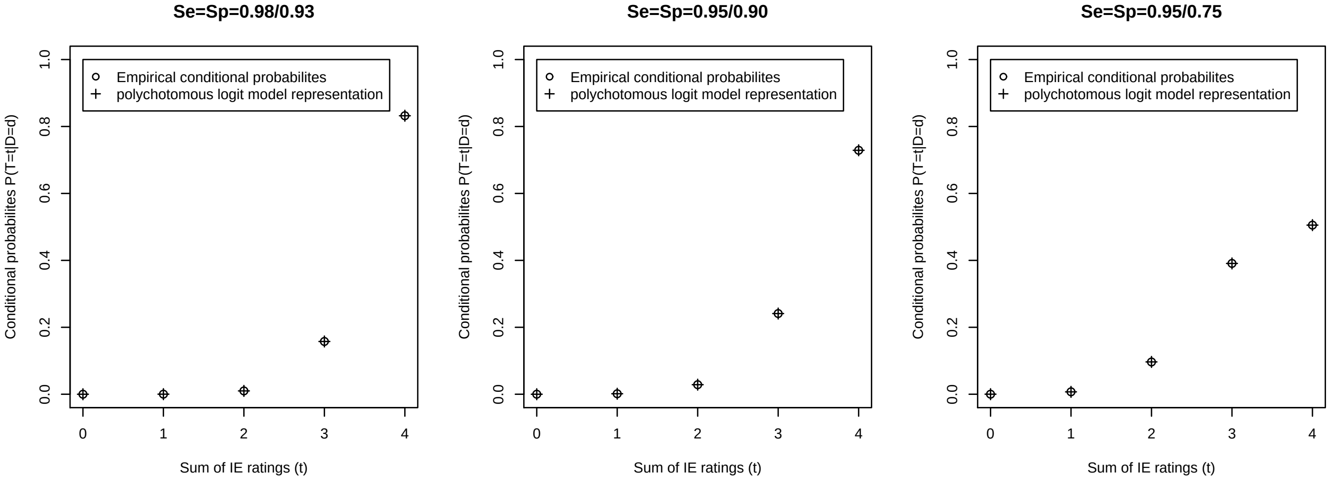

Figure S.2: Empirical conditional probabilities and their polychotomous logit model representations when (1) two of the four IEs had sensitivity and specificity 0.98 and another two had 0.93, (2) two had sensitivity and specificity 0.95 and another two had 0.90, and (3) two had sensitivity and specificity 0.95 and another two had 0.75.

Hypothetically, if the IEs are not exchangeable (i.e., $P(\tilde{T}_i^{(1)} = e_1, \tilde{T}_i^{(2)} = e_2, \tilde{T}_i^{(3)} = e_3, \tilde{T}_i^{(4)} = e_4 | D_i = d_i)$ varies with the different values of (e_1, e_2, e_3, e_4) and does not only depend on $\sum_{l=1}^{4} e_l$), the approach might lead to biased estimation. However, we show that, in the scenario of non-exchangeability, the polychotomous logit model is still able to capture the conditional probabilities of the sum of the IE ratings $S^T_{t_i|d_i}$. Further, and more importantly, the estimates of the sensitivity and specificity of the R-OB/GYNs are unbiased when the exchangeable assumption is violated. The following discussion consists of two parts: first, to show the flexibility of polychotomous logit model (Equation (7) in the manuscript), we demonstrate through a large simulated dataset that the polychotomous logit model is able to form an appropriate representation for the conditional probabilities of the sum of the

IE ratings given the true disease status in the scenario of non-exchangeability; second, the simulation study in Table 1(B) was similarly conducted with four non-exchangeable IEs and the results show that the estimates of the sensitivity and specificity of the R-OB/GYNs remain nearly unbiased.

Part I: Flexibility of polychotomous logit model on non-exchangeability

We simulated a large dataset with 50000 observations, where each observation contained the ratings from four IEs. Under the conditional independence assumption (although this works more generally), we assumed two of the four IEs had larger sensitivity and specificity than the other two IEs. Based on the large dataset, we were able to show that the polychotomous logit model characterized $S^T_{t_i|d_i}$ remarkably well under non-exchangeability. We fit the polychotomous logit model to the simulated data and estimated the model parameter of the polychotomous logit regression (Equation (7) in the manuscript) as their model representations. Figure S.2 shows the empirical conditional probabilities and their polychotomous logit model representations when (1) two of the four IEs had sensitivity and specificity 0.98 and another two had 0.93, (2) two had sensitivity and specificity 0.95 and another two had 0.90, and (3) two had sensitivity and specificity 0.95 and another two had 0.75. From Figure 1, we can conclude that the polychotomous logit model describes the correct conditional distribution of the sum of the IEs, even when the four IEs are not exchangeable. We also examined other cases where the four IEs had different combination of sensitivity and specificity. In all cases, the polychotomous logit model did a very good job representing $S^T_{t_i|d_i}$.

Part II: Robustness of sensitivity and specificity estimation

To investigate the robustness of the estimates of the sensitivity and specificity of the R-OB/GYNs when the IEs are non-exchangeable, we conducted a simulation study similar to Table 1 (B) in the manuscript. The ratings of the IEs were generated by assuming two of the four IEs had sensitivity and specificity of 0.98 and another two had 0.93. As in Table 1(B), the parameters of the polychotomous logit model characterizing $S^T_{t_i|d_i}$ are assumed to be known. Table S.4 in this document shows the simulation results. The estimated sensitivity and specificity of R-OB/GYNs are nearly unbiased under the correctly specified conditional probabilities $S^T_{t_i|d_i}$.

Table S.4: Simulation results for sensitivity and specificity under the estimated imperfect reference standard from the polychotomous logit model when the IEs are nonexchangeable. The random effects of the true models follow mixture normal (MixN) distribution. The averages of estimates (standard errors) and the percentage of selecting true model by $\text{IC}_{H(0),Q}$ are presented. The true sensitivity, specificity and disease prevalence are $S_e = 0.88$, $S_p = 0.87$, and $\pi_1 = 0.7$, respectively.

Number of tests	Working random effects distribution	$\hat{S}_e(se)$	$\hat{S}_p(se)$	$\hat{\pi}_1(se)$	Rate of selecting true model
5	Normal	0.88(0.063)	0.87(0.056)	0.70(0.053)	92%
	MixN	0.88(0.061)	0.87(0.064)	0.70(0.058)	
10	Normal	0.87(0.055)	0.87(0.054)	0.70(0.052)	95%
	MixN	0.88(0.059)	0.87(0.051)	0.70(0.049)	

Although the approach is robust to non-exchangeability among the IEs, we have reasons to assume exchangeability of the IEs for the PRS. The IEs in the PRS are selected by the investigators to be well known international experts with an equivalent amount of expertise.

Web Appendix F. On the violation of the independence of random effect b_i and the imperfect reference T_i

In the article, we assume the independence of random effect b_i and the imperfect reference T_i. Now we further investigate features and performance of the proposed methodology when the assumption is violated. It is reasonable to assume that, conditional on the disease status, subjects who are diagnosed as diseased are more likely to be diagnosed as diseased by the IEs. In this circumstance, the assumption of independence between the random effect b_i and the imperfect reference standard T_i is violated. To link the ratings from the IEs and the ratings from the R-OB/GYNs, it is natural to incorporate random effects into the polychotomous logit model. As a result, we consider the following random-effects polychotomous logit model

$$P(T_i = t_i|D_i = 1, b_i) = \frac{\exp(\gamma_0 + \gamma_1 t_i + \gamma_2 t_i^2 + \tau b_i)}{1 + \sum_{h=0}^{3} \exp(\gamma_0 + \gamma_1 h + \gamma_2 h^2 + \tau b_i)}, \qquad t_i = 0, 1, \cdots, 3, \qquad (3)$$

where the random effect b_i is shared with Equation (5) in the manuscript. We now examine the features of the polychotomous logit model and the robustness of the estimates of the sensitivity and specificity of the R-OB/GYNs when the assumption of independence between the random effect b_i and the imperfect reference standard T_i is violated. Our discussion consists of two parts: first, we show that the polychotomous logit model still provides a good representation of the conditional probabilities of the sum of the IE ratings $S^T_{t_i|d_i}$ when the independence assumption is violated; second, we show that ignoring the conditional dependence between b_i and T_i results in nearly unbiased estimates of the sensitivity and specificity of the R-OB/GYNs and unbiased estimates of the prevalence of endometriosis.

Part I: Flexibility of polychotomous logit model on non-exchangeability

We generated a large simulated dataset with 50000 observations from the random-effects polychotomous logit model (3), where each observation contained the ratings from four IEs. Figure S.3 shows the empirical conditional probabilities and the polychotomous logit model representations (1) when $\tau = -0.5$ and (2) $\tau = -0.25$. Here, we fit the polychotomous logit model without random effect (Equation (7) in the manuscript) to the data, and obtaining the parameters of the model corresponding to $S^T_{t_i|d_i}$. From Figure S.3, we can conclude that the polychotomous logit model (Equation (7) in the manuscript) nicely characterizes $S^T_{t_i|d_i}$ when the conditional dependence is ignored. We examined other values of τ and obtained similar results.

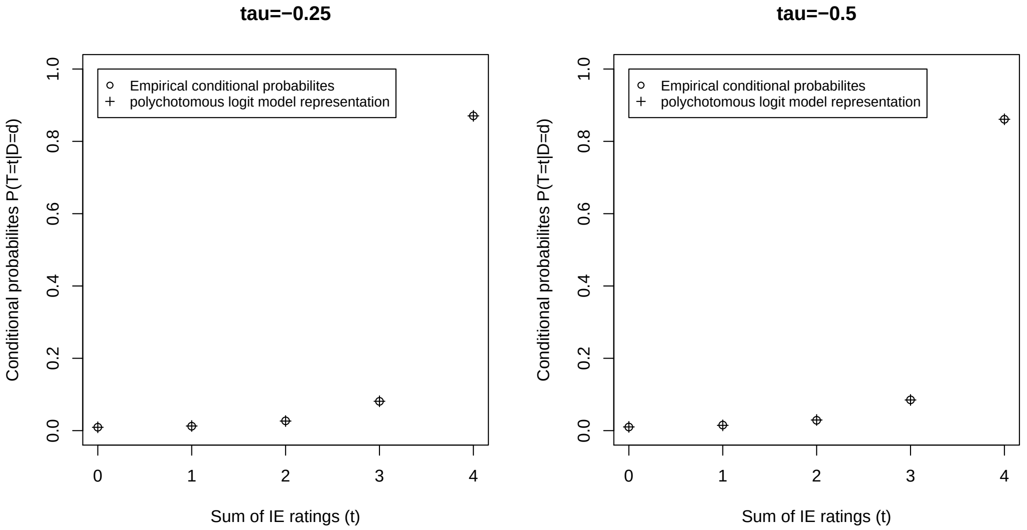

Figure S.3: Empirical conditional probabilities and their polychotomous logit model representations when $\tau = -0.25$ and $\tau = -0.5$.

Part II: Robustness of sensitivity and specificity estimation

We further investigate the robustness of the estimates of the sensitivity and specificity of the R-OB/GYNs when the IE ratings come from (3) with $\tau = -0.5$. The simulation study in Table 1 (B) in the manuscript was repeated with the ratings of the IEs generated from (3). As in the standard approach proposed in the manuscript, the polychotomous logit model parameters that characterize $S^{T}_{t_i|d_i}$ are assumed to be known as the imperfect reference standard. Table S.5 in this document shows the simulation results. The estimated sensitivity and specificity of R-OB/GYNs are nearly unbiased even when the assumption of independence between the random effect b_i and the imperfect reference standard T_i is violated.

Table S.5: Simulation results for sensitivity and specificity under the estimated imperfect reference standard from the polychotomous logit model when the IEs ratings are simulated from the random-effects polychotomous logit model (3). The random effects of the true models follow mixture normal (MixN) distribution. The averages of estimates (standard errors) and the percentage of selecting true model by $\mathrm{IC}_{H(0),Q}$ are presented. The true sensitivity, specificity and disease prevalence are $S_e = 0.88$, $S_p = 0.87$, and $\pi_1 = 0.7$, respectively.

Number of tests	Working random effects distribution	$\hat{S}_e(se)$	$\hat{S}_p(se)$	$\hat{\pi}_1(se)$	Rate of selecting true model
5	Normal	0.88(0.057)	0.86(0.053)	0.70(0.058)	92%
	MixN	0.88(0.063)	0.87(0.063)	0.70(0.056)	
10	Normal	0.88(0.051)	0.87(0.061)	0.69(0.054)	93%
	MixN	0.88(0.053)	0.86(0.065)	0.70(0.059)	

References

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **61**, 265–285.

Fenton, V. M., and Gallant, A. R. (1996). Qualitative and Asymptotic Performance of SNP Density Estimators. *Journal of Econometrics*, **74**, 77–118.

Gallant, A. R., and Nychka, D. W. (1987). Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica* **55**, 363–390.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* New York: Chapman and Hall/CRC.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model Selection Criteria for Missing-Data Problems Using the EM Algorithm. *Journal of the American Statistical Association* **103**, 1648–1658.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.