Simulation studies on the estimation of sensitivity and specificity in the absence of a gold standard and with rare positive frequencies

Authors

Tasuku Inao[1], Kazufumi Okada[1], Yichi Yang[1], Isao Yokota[1]

1. Department of Biostatistics, Graduate School of Medicine, Hokkaido University, Sapporo, Japan

Correspondence

Isao Yokota, Department of Biostatistics, Graduate School of Medicine, Hokkaido University, N15, W7, Kita-ku, Sapporo, Hokkaido, Japan

e-mail: yokotai@pop.med.hokudai.ac.jp

Funding Information

Health, Labour and Welfare Policy Research Grants 20HA2002, and AMED under Grant Number JP20fk0108471, JP21fk0108489.

Abstract

In the absence of a gold standard, sensitivity and specificity could be evaluated using the Bayesian latent class model underlying true positives and true negatives. However, when the frequency of positivity is limited owing to low prevalence, the estimates may be strongly influenced by prior information or small-sample bias. In this study, we evaluated the performance of the Bayesian latent class model with two independent tests under a rare frequency of positivity, varying the strength and way of giving prior information under the conditional independence assumption is satisfied. Throughout the simulation experiments, a small-sample bias led to underestimation and overestimation when the frequency of positivity was low. Furthermore, we observed that placing an informative prior distribution for only one of the two tests resulted in a constant bias of sensitivity or specificity for the other test, regardless of the strength of the prior information.

Keywords: Diagnostic test, finite population, latent class model, Markov chain Monte Carlo, sparse data

1 Introduction

To prevent the spread of infectious diseases, such as coronavirus diseases 2019 (COVID-19), it is essential to develop diagnostic methods that detect transmissive individuals who should be quarantined. The gold standard is whether a transmissive person should be quarantined, and such an ideal test has not been developed yet. Diagnostic performance should be examined under conditions that allow for less than perfect sensitivity and specificity. Traditionally, we test repeatedly in a particular cohort, and the sensitivity and specificity are calculated by whether the future disease develops or no as the gold standard. Because of the time lag between exposure and onset of disease, it is often difficult to determine whether the patient was infected at the time of each test or whether they were transmissive.

On obtaining multiple binary test results, the sensitivity and specificity can be estimated based on the latent class model. In an earlier report[1,2], the latent model was imposed on the conditional independence (CI) assumption that the diagnosis was independent across tests conditional on the latent class. The latent class model can lead to bias in the estimates by violating the CI assumption[3,4] and heterogeneity in sensitivity or specificity within a population[5]. In particular, Pepe et al.[6–8] have criticized the CI assumption not only for its strength, but also that it cannot be fully verified whether the assumption holds and that the latent class does not provide a clinical definition of the disease. In terms of the CI assumption, several methods have been proposed to relax the assumption using more than two latent classes[9], random effects[10], and hierarchical models[11–13]. Nevertheless, latent class models are still characterized by their sensitivity to misspecification[14].

The identifiability of the latent class model has been discussed based on Jacobian analysis[15,16]. This indicates that the latent class model for the two tests is not identifiable regardless of the CI assumption or data from multiple populations, and conventional maximum likelihood estimation cannot be adapted. In contrast, the Bayesian method may enable estimation using the prior of parameters effectively. In addition, we can impose the constraint on the range of probability for each parameter. Previously, many issues have been pointed out for the Bayesian latent class model[17,18].

Mass screening of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) na-

sopharyngeal swabs and patient saliva simultaneously was carried out and the CI model was applied using latent classes for reverse transcription-polymerase chain reaction of each specimen[19]. The screening showed excellent sensitivity and near-perfect specificity of both nasopharyngeal-based and saliva-based tests. When the prevalence is low, such as in the mass screening of the general population, the frequency of positivity is rare. To deal with sparseness, the use of the Jeffreys-prior penalty[20,21] or weak prior[22–24] has been proposed. Most often, the performance is known for some diagnostics. Performing Bayesian estimation using this information or beliefs as a prior distribution would be useful. However, because of sparse data, the amount of prior information has a relatively large impact on the parameter estimation. Therefore, in this study, we aimed to investigate the impact of the strength of the prior distribution on the estimation of sensitivity and specificity in latent class models in situations where (1) two independent tests were applied for one population and (2) the frequency of positive results was rare.

2 Latent class model

2.1 Conditional independence model

Suppose that all subjects have positive or negative results for tests 1 and 2, respectively. Let $y_{ij}(i, j = 0(\text{negative}), 1(\text{positive}))$ be the frequency of the diagnosis pattern, as shown in Table 1(a). We assumed that two latent classes are reflected in the presence ($D = 1$) or absence ($D = 0$) of the infection. Note that the true infection here is defined in the latent class model and cannot be observed. As shown in Table 1(b), in the infected latent class, the sensitivity in each test is observed as positive and (1-sensitivity) is observed as a false positive. Because the two tests are assumed to give results conditionally independent of latent classes, the diagnostic pattern is expressed as the product of the sensitivities of each test. A similar probability model is expressed using specificity in the non-infected latent class.

[Table 1 about here.]

In the CI latent class model described above, the probability π of observing a pattern of testing is expressed as follows:

$$\pi_{11} = pSe1Se2 + (1-p)(1-Sp1)(1-Sp2)$$
$$\pi_{10} = pSe1(1-Se2) + (1-p)(1-Sp1)Sp2$$
$$\pi_{01} = p(1-Se1)Se2 + (1-p)Sp1(1-Sp2)$$
$$\pi_{00} = p(1-Se1)(1-Se2) + (1-p)Sp1Sp2$$

where p denotes the prevalence. Because of the independence between each combination for the diagnostic by two tests, a quadrinomial distribution is applied to the likelihood. Then, the log-likelihood function is determined using the following equation:

$$\log L = y_{11}\log(\pi_{11}) + y_{10}\log(\pi_{10}) + y_{01}\log(\pi_{01}) + y_{00}\log(\pi_{00}).$$

2.2 Bayesian estimation

The parameters are unidentifiable in the standard maximum likelihood estimation because the number of parameters exceeds the number of degrees of freedom in this probabilistic model. Alternatively, the parameters are estimated using the Bayesian approach. We used the beta distribution as prior information as doing so has the following two advantages: (1) beta distribution can express any probability by changing arguments, and (2) the posterior distribution is also a beta distribution because the beta distribution is a conjugate prior distribution for likelihood. Bayesian estimation with some prior information[2] generates the joint posterior distribution of parameters using Markov chain Monte Carlo (MCMC).

Jones et al.[16] reported that it is reasonable to constrain the parameter space for the sensitivity and specificity in the same test such that the sum of the two is greater than one because of label switching. Therefore, we addressed this problem by specifying the log-likelihood function as -1e+200, if it is not $Se1 \geq 1 - Sp1$ or $Se2 \geq 1 - Sp2$ in the MCMC sampling.

Because of the large number of parameters and the expected skewed distribution of specificity, a blocking Metropolis-Hastings algorithm was adopted to improve the efficiency

of MCMC[25]. Specifically, we used the three-block strategy for the MCMC sample as $\{Se1, Se2\}, \{Sp1, Sp2\}, \{p\}$.

3 Simulation study

3.1 Design

3.1.1 Simulation 1: The performance of estimation with the strength of prior information

We investigated a situation in which the positivity frequency was low due to mass screening in a low prevalence population. In this case, most individuals would have negative results for both, and the frequency of one or both being positive would be rare. In addition to the small-sample bias due to sparse data, the posterior distribution should be relatively strongly influenced by the prior distribution because it is a shrinkage estimator. We carried out Monte Carlo simulations under several scenarios with different sets of prior distributions.

Because all parameters are binomial proportions, prior and data-generating distributions were set based on beta and binomial distributions. We denote the prior effective sample size as N_{ESS}, which expresses the strength of the prior distribution[26,27]. Since we aimed to assess the bias in finite samples, we set the expectation of prior distributions to the true values of sensitivity η, specificity ζ, and prevalence θ, except for the uninformative prior. In the case of an uninformative prior, the distribution was $Uniform(0, 1)$, that is, a uniform distribution.

[Table 2 about here.]

Table 2 provides a summary of the scenarios. We examined situations in which both sensitivity and specificity were equal between the two tests. We tried four such scenarios mainly. Scenario 1 had informative prior distributions for sensitivity and was divided into two scenarios: scenario 1-1 and scenario 1-2. Scenario 2 had informative prior distributions for specificity and was also divided into two scenarios: scenario 2-1 and scenario 2-2. In addition, in both scenarios 1 and 2, we set up the situations in which we had an informative

prior for both tests and only one of the tests. Scenario 3 had informative prior distributions for sensitivity and specificity for only one of the tests. Finally, scenario 4 had informative prior distributions for the prevalence. We also worked out scenario 0 with uninformative prior distributions for all parameters.

The strength of the prior information, N_{ESS} was varied as 5, 10, 20, 30, 50, 100, 200, 300, and 500. As a reference result, we also evaluated the situation where the informative prior was fixed by a degenerate distribution at the true value, which corresponds to N_{ESS} as infinity. We set the sample size to 2,000, the true sensitivity η to 0.7, and the true specificity ζ to 0.99. The true prevalence θ was varied as 0.1%, 0.5%, 1.0%, and 10%. The point estimates were defined as the mean of the posterior distribution generated using 50,000 MCMC samples after 5,000 burn-in periods without thinning. We evaluated the bias, statistical efficiency, and convergence of MCMC using the absolute bias, root mean squared error (RMSE), and the probability of autocorrelation with 500 lags being less than 0.1 for 10,000 repetitions. All simulations were implemented using proc mcmc in SAS®9.4.

3.1.2 Simulation 2: Varying the true value of sensitivity

We assessed the absolute bias and RMSE by changing the true value of sensitivity, where η was set to 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. The informative prior in each scenario was fixed using a degenerate distribution at the true value. ζ and θ were set to the same values as in simulation 1, and the settings of the Monte Carlo simulation were the same as simulation 1.

3.1.3 Simulation 3: Comparison with existing frequentist methods

We compared the Bayesian method with original Hui and Walter[1], and Lu *et al.*[28] that modified the likelihood function to increase the stability in estimation. Under holding the CI assumption, the sensitivities and specificities were estimated in two populations of size 1,800 and 200. The true sensitivity η was set to 0.5, 0.6, 0.7, 0.8, and 0.9, the true specificity ζ was set to 0.99, and the prevalence θ was set to 0.5% and 10%. In Bayesian estimation, the prior of specificity, $Sp1$ and $Sp2$, was informative with $N_{ESS} = 200$ and

the other parameter of sensitivity and prevalence was uninformative.

3.2 Simulation results

3.2.1 Simulation 1: The performance of estimation with the strength of prior information

We present some representative results in this section; all the results are presented in the supplementary material. Regarding sensitivity, the performances for $Se1$ and $Se2$ were similar in scenarios 1-1, 2-1, and 4, where the prior distributions were defined symmetrically with each other. The results for $Se1$ were similar between scenarios 1-1 and 1-2, with the same informative prior for $Se1$. Scenario 2-1 had informative priors for both specificity and scenario 2-2 for Sp_1. This shows that the bias in $Se2$ decreases as the prior for specificity becomes stronger. Although $Se1$ shows a smaller bias in scenario 2-1 as the prior becomes stronger, it shows a constant bias in scenario 2-2 regardless of the strength of the prior distribution for $Sp1$.

Figure 1 shows the simulation results for $Se2$ with a prevalence of 0.5% and 10%. The convergence rate in scenarios 1-1, 1-2, and 4 was almost 1.0 in all situations. On the contrary, this rate in scenarios 2-1, 2-2, and 3 was less than 1.0 when the effective sample size was smaller than 100. Moreover, the rate in scenario 2-1, when effective sample size is from 10 to 20, was lower compared with that when the size was 5. In scenario 1-2, in which the only Se2 has an informative prior, we observed a constant bias, irrespective of how strongly the information was fed in. The bias in scenario 2-1 and 2-2 was almost the same value by same effective sample size. The bias in scenario 3 was also almost the same value by same effective sample size compared with the bias in scenario 2-1 and 2-2; however, the absolute bias in scenario 3 on a fixed effective sample size was larger than the bias in scenarios 2-1 and 2-2. As long as Se2 was a uninformative prior (except for in scenario 1-1), a large absolute bias of more than 10% was observed in lower prevalence situations, namely, low-positive frequencies. In scenario 1-1, the bias moved toward zero as the prior distribution strengthened, despite the low prevalence. In addition, the bias generally moved toward zero as the prior distribution information increased at a higher

prevalence.

[Figure 1 about here.]

Figure 2 shows the simulation results for $Sp2$ with a prevalence rate of 0.5% and 10%. The convergence rate in scenarios 1-1, 1-2, and 4 was almost fair in all situations. This rate in scenarios 2-1, 2-2, and 3 increased as the effective sample size increased, and the convergence ratio was almost 1.0 when the effective sample size was greater than 100. In scenario 2-2, in which the only Sp2 has an informative prior, we noticed a constant bias, regardless of how strongly we put the information. However, the bias in all scenarios, except for scenario 2-2, was asymptotically 0 as the effective sample size increased. In all situations, absolute bias observed higher prevalence rate is larger than lower.

Because the negative frequency was relatively high, it was found that the bias could be generally reduced if some prior information was provided. Uniquely, when estimating specificity, bias was smaller, and the convergence rate was better in scenario 1 with informative prior for sensitivity than in scenario 2, which has an informative prior distribution. As with the above considerations for sensitivity, when an informative prior distribution is placed on the other test exclusively, a constant bias regardless of the strength of the information is inevitable.

[Figure 2 about here.]

3.2.2 Simulation 2: Varying the true value of sensitivity

Figure 3 shows the results of Se2 bias and RMSE at prevalence rates of 0.5% and 10% for variations in the true value of sensitivity. In scenario 1-2, the absolute bias was large compared with that in other scenarios where the prevalence was 0.5%. In situations where positive frequencies were rare, sensitivity was underestimated in all settings. In particular, scenario 1-2, in which only $Se1$ was fixed at its true value, had a severe bias compared with that observed in the previous method. In situations where the prevalence was 10%, although the bias was generally small given an informative prior in specificity (scenario 2) or prevalence (scenario 4), scenario 1-2 showed bias in both directions depending on the true sensitivity.

[Figure 3 about here.]

3.2.3 Simulation 3: Comparison with existing frequentist methods

In the Hui-Walter method, there were cases where parameters were estimated outside the range of 0 to 1. Also in the Lu's modified likelihood method, there were cases where the likelihood was maximized when any of the parameter took 0 or 1. In this study, we regarded these cases as inestimable. The performances of bias, RMSE, and estimable proportion in the sensitivity Se_2 are shown in Figure 4.

[Figure 4 about here.]

Jacobian analysis already revealed that the situation of two-test and two-population was unidentifiable and the Hui-Walter method could not be estimated in most cases. For the 0.5% prevalence setting, the bias of the Hui method appears to smaller than the other methods, but it summarizes only the few repetitions that could be estimated. Comparing the case with a prevalence of 0.5% to the case with a prevalence of 10%, the bias of the Bayesian method was smaller for the 10% prevalence, where the frequency of positive cases is higher. The Lu's method also showed a similar trend of bias, still with some cases that could not be estimated. Throughout, the RMSE of the Bayesian method was smaller than the RMSE of the other methods implying that the estimation is more stable.

4 Discussion

In this study, we assessed the Bayesian estimates of sensitivity and specificity from the two test results using a CI model with latent classes. In the absence of a gold standard, the performance of a convenient statistical method for finite samples for developing and evaluating clinical laboratory testing has been demonstrated. We found a certain magnitude of bias of the sensitivity and specificity in finite samples even if the CI assumption is held.

Five essential points can be listed based on the results of Se2 in simulation 1. First, sampling by the MCMC method is inefficient because of the low convergence rate in scenario 2-1, 2-2, and 3 when the effective sample size is lower than 50. Therefore, we recommend

using a sample size of 100 or more to improve the bias. Second, the absolute bias is reduced by giving informative prior to Sp1 based on the results obtained on comparing 1-2 with 3. We can suggest that, although the bias of Se2 is not improved regardless of the effective sample size, the sensitivity of the other test (Se1) increases. Third, absolute bias remains even if the effective sample size is fixed at the true value in scenarios, except for 0 and 1-1. Fourth, the bias of scenarios 2-1, 2-2, and 3 in the effective sample size that is lower than 500 is almost the same for each size, so the scenario providing information to only Sp1 can be used to represent the change of bias for both the scenarios providing information to Sp1 and Sp2, and the scenario providing information to Sp1 and Se1. Finally, the absolute bias of Se2 in low prevalence is larger than that in high prevalence for each scenario, so estimating the sensitivity is highly influenced and underestimated owing to the bias of Se2 in diseases where the frequency of positive results is rare, such as that observed in this study.

Four essential points can be listed based on the results of Sp2 in simulation 1. First, sampling by the MCMC method is inefficient because of the low convergence rate in scenarios 2-1, 2-2, and 3 when the effective sample size is lower than 50, similar to the results in Se2 of simulation 1. Second, the bias in scenario 2-1 is larger than that in scenario 2-2 for an effective sample size of less than 200. We found that giving informative prior to Sp1 and Sp2 is worse than giving informative prior to only Sp1 for the bias of Sp2 when the effective sample size is less than 200. Third, because the change in bias in scenario 2-2 does not affect the magnitude of the effective sample size, and the change in bias in scenarios 1-1, 1-2, and 3 is the same, we can say that the estimation of Sp2 is highly influenced by the informative prior given to Se1. Finally, the absolute bias of Sp2 in low prevalence is smaller than that in high prevalence for each scenario, so the bias of Sp2 is not an important factor for consideration when estimating the specificity in diseases for which the frequency of positive results is rare, such as that observed in this study.

Compared to the frequentist methods shown in simulation 3, the Bayesian method tends to have more stable estimates and smaller bias. This is because it can be computed as a shrinkage estimator by giving a prior distribution. In the practical development of diagnos-

tic tests, information on sensitivity and specificity of the conventional test or measurement of specificity using negative controls will be performed, so a Bayesian method that can utilize such information would be valuable.

Two novel essential points can be considered based on the simulation studies. When a low frequency of positivity is observed, there is a small-sample bias of the sensitivity estimates in both directions, unless there is a strong prior distribution for the parameter directly. Although there is a tendency for the bias for sensitivity to be smaller when a strong prior distribution is placed in specificity or prevalence, it remains a certain bias even when the degenerate distribution at the true value is given.

Second, we found that giving an informative prior distribution for sensitivity to only one test leads to severe bias in estimating the sensitivity of other tests. A similar trend in specificity was observed in situations where a modest negative frequency was observed. In scenario 1-2, the bias of $Se1$ and prevalence improved as the prior distribution for $Se1$ was strengthened, while the bias of $Se2$ did not improve. This leads us to think that the information obtained from observational data may contribute to the estimation of parameters for which we already have prior information. However, since scenarios 1-2 and 2-2 have four parameters with uninformed prior distributions, which is larger than the observed data with 3 degrees of freedom, the general validity of this argument is doubtful. Nevertheless, we recommend that an informative prior distribution for either sensitivity or specificity should be provided in both tests.

There is a good convergence status of MCMC. In our simulation study, the three-block strategy significantly improved the autocorrelation compared to that by simultaneous MCMC sampling. Scenario 2 presented some degree of autocorrelation in the MCMC-derived samples. Because of the rarity of positive frequencies, the MCMC series would likely be more stable if we placed an informative prior distribution on the less informative sensitivities and prevalence provided by the data.

In our simulations, we only evaluated limited situations in which we placed an informative prior distribution on only one or two of the five estimated parameters. For three or more tests, it is possible to fit a conditional dependence model with latent classes,

which incorporates the correlation between tests into the model. Our study suggests that a small-sample bias may be present even in these complex models. Furthermore, if the CI assumption violates, the bias could be greater and further evaluation in finite sample would be valuable.

Finally, in developing new diagnostic methods, it is essential to compare their diagnostic performance with that of previously established diagnostic methods. Because the concordance probability between the two tests is a measure depending on the prevalence, it is crucial to decompose the results into sensitivity and specificity and to discuss the estimated concordance probability under various prevalence rates. However, we must be cautious about the impact of CI assumptions in interpreting the estimated sensitivity and specificity[29], and clinical interpretations should be considered for the estimated latent classes.

Acknowledgments

We thank Dr. Fang Chen from SAS Institute Inc., Yusuke Ono from SAS Institute Japan, and Yoko Unoki from Hokkaido University for their helpful comments. We would also like to thank the editorial team and the anonymous reviewer for helpful comments on an earlier version of this manuscript.

IY reports grants from KAKENHI, AMED, and Health, Labour and Welfare Policy Research Grants, research fund by Nihon Medi-Physics, and speaker fees from Chugai Pharmaceutical Co, AstraZeneca plt, and Nippon Shinyaku Co, outside the submitted work. All other authors report no potential conflicts.

Data availability statement

The motivated example data are distributed in the supplement file of the original article[19].

References

[1] Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics* 1980; 36(1): 167–171.

[2] Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol* 1995; 141(3): 263–272. http://dx.doi.org/10.1093/oxfordjournals.aje.a117428 doi: 10.1093/oxfordjournals.aje.a117428

[3] Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 1985; 41(4): 959–968.

[4] Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; 60(2): 427–435.

[5] Toft N, Jørgensen E, Højsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. *Prev Vet Med* 2005; 68(1): 19-33. http://dx.doi.org/10.1016/j.prevetmed.2005.01.006 doi: 10.1016/j.prevetmed.2005.01.006

[6] Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007; 8(2): 474-484. http://dx.doi.org/10.1093/biostatistics/kxl038 doi: 10.1093/biostatistics/kxl038

[7] Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Stat Med* 1999; 18(22): 2987–3003.

[8] Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press . 2003.

[9] Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Stat Med* 1986; 5(1): 21-27. http://dx.doi.org/10.1002/sim.4780050105 doi: 10.1002/sim.4780050105

[10] Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 1996; 52(3): 797–810.

[11] Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; 57(1): 158–167. http://dx.doi.org/10.1111/j.0006-341X.2001.00158.x doi: 10.1111/j.0006-341X.2001.00158.x

[12] Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: A multiple latent variable model. *Stat Med* 2009; 28(3): 441-461. http://dx.doi.org/10.1002/sim.3470 doi: 10.1002/sim.3470

[13] Wang C, Lin X, Nelson KP. Bayesian hierarchical latent class models for estimating diagnostic accuracy. *Stat Methods Med Res* 2020; 29(4): 1112–1128. http://dx.doi.org/10.1177/0962280219852649 doi: 10.1177/0962280219852649

[14] Schofield MR, Maze MJ, Crump JA, Rubach MP, Galloway R, Sharples KJ. On the robustness of latent class models for diagnostic testing with no gold standard. *Stat Med* 2021; 40(22): 4751-4763.

[15] Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; 61(2): 215-231. http://dx.doi.org/10.1093/biomet/61.2.215 doi: 10.1093/biomet/61.2.215

[16] Jones G, Johnson WO, Hanson TE, Christensen R. Identifiability of Models for Multiple Diagnostic Testing in the Absence of a Gold Standard. *Biometrics* 2010; 66(3): 855–863. http://dx.doi.org/10.1111/j.1541-0420.2009.01330.x doi: 10.1111/j.1541-0420.2009.01330.x

[17] Collins J, Huynh M. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine* 2014; 33(24): 4141-4169. http://dx.doi.org/https://doi.org/10.1002/sim.6218 doi: https://doi.org/10.1002/sim.6218

[18] Kostoulas P, Nielsen SS, Branscum AJ, et al. STARD-BLCM: standards for the reporting of diagnostic accuracy studies that use Bayesian latent class models. *Preventive Veterinary Medicine* 2017; 138: 37-47. http://dx.doi.org/https://doi.org/10.1016/j.prevetmed.2017.01.006 doi: https://doi.org/10.1016/j.prevetmed.2017.01.006

[19] Yokota I, Shane PY, Okada K, et al. Mass screening of asymptomatic persons for SARS-CoV-2 using saliva. *Clin Infect Dis* 2020. http://dx.doi.org/10.1093/cid/ciaa1388 doi: 10.1093/cid/ciaa1388

[20] Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; 80(1): 27-38. http://dx.doi.org/10.1093/biomet/80.1.27 doi: 10.1093/biomet/80.1.27

[21] Kosmidis I, Firth D. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* 2020. [Epub ahead of print]http://dx.doi.org/10.1093/biomet/asaa052 doi: 10.1093/biomet/asaa052

[22] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat* 2008; 2(4): 1360–1383. http://dx.doi.org/10.1214/08-AOAS191 doi: 10.1214/08-AOAS191

[23] Hamra GB, MacLehose RF, Cole SR. Sensitivity analyses for sparse-data problems—using weakly informative Bayesian priors. *Epidemiology* 2013; 24(2): 233–239. http://dx.doi.org/10.1097/EDE.0b013e318280db1d doi: 10.1097/EDE.0b013e318280db1d

[24] Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med* 2015; 34(23): 3133–3143. http://dx.doi.org/10.1002/sim.6537 doi: 10.1002/sim.6537

[25] Roberts GO, Sahu SK. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J R Stat Soc Series B Stat Methodol* 1997; 59(2): 291–317. http://dx.doi.org/10.1111/1467-9868.00070 doi: 10.1111/1467-9868.00070

[26] Morita S, Thall PF, Müller P. Determining the Effective Sample Size of a Parametric Prior. *Biometrics* 2008; 64(2): 595-602. http://dx.doi.org/10.1111/j.1541-0420.2007.00888.x doi: 10.1111/j.1541-0420.2007.00888.x

[27] Morita S, Thall PF, Müller P. Prior effective sample size in conditionally independent hierarchical models. *Bayesian Anal* 2012; 7(3): 591–614. http://dx.doi.org/10.1214/12-BA720 doi: 10.1214/12-BA720

[28] Lu D, Zhou C, Tang L, Tan M, Yuan A, Chan L. Evaluating accuracy of diagnostic tests without conditional independence assumption. *Statistics in Medicine* 2018; 37(19): 2809-2821.

[29] Smeden vM, Naaktgeboren CA, Reitsma JB, Moons KGM, Groot dJAH. Latent class models in diagnostic studies when there is no reference standard—a systematic review. *Am J Epidemiol* 2013; 179(4): 423–431. http://dx.doi.org/10.1093/aje/kwt286 doi: 10.1093/aje/kwt286

List of Tables

Table 1: Observed patterns and probabilistic model for two tests.

(a) Observed frequencies.

	test 2	
test 1	positive	negative
positive	y_{11}	y_{10}
negative	y_{01}	y_{00}

(b) Probabilistic model for conditional independent model.

Infectious status (latent class)					
Infected, $D = 1$			Non-infected, $D = 0$		
	test 2			test 2	
test 1	positive	negative	test 1	positive	negative
positive	$Se_1 \cdot Se_2$	$Se_1(1 - Se_2)$	positive	$(1 - Sp_1)(1 - Sp_2)$	$(1 - Sp_1)Sp_2$
negative	$(1 - Se_1)Se_2$	$(1 - Se_1)(1 - Se_2)$	negative	$Sp_1(1 - Sp_2)$	$Sp_1 \cdot Sp_2$

Table 2: Details of prior distribution in each scenario.

Scenario	Informative prior	Uninformative prior
0		$Se1, Se2, Sp1, Sp2, p \sim Uniform(0,1)$
1-1	$Se1 \sim Beta(\eta N_{ESS}, (1-\eta)N_{ESS})$ $Se2 \sim Beta(\eta N_{ESS}, (1-\eta)N_{ESS})$	$Sp1, Sp2, p \sim Uniform(0,1)$
1-2	$Se1 \sim Beta(\eta N_{ESS}, (1-\eta)N_{ESS})$	$Se2 \sim Uniform(0,1)$ $Sp1, Sp2, p \sim Uniform(0,1)$
2-1	$Sp1 \sim Beta(\zeta N_{ESS}, (1-\zeta)N_{ESS})$ $Sp2 \sim Beta(\zeta N_{ESS}, (1-\zeta)N_{ESS})$	$Se1, Se2, p \sim Uniform(0,1)$
2-2	$Sp1 \sim Beta(\zeta N_{ESS}, (1-\zeta)N_{ESS})$	$Sp2 \sim Uniform(0,1)$ $Se1, Se2, p \sim Uniform(0,1)$
3	$Se1 \sim Beta(\eta N_{ESS}, (1-\eta)N_{ESS})$ $Sp1 \sim Beta(\zeta N_{ESS}, (1-\zeta)N_{ESS})$	$Se2 \sim Uniform(0,1)$ $Sp2 \sim Uniform(0,1)$ $p \sim Uniform(0,1)$
4	$p \sim Beta(\theta N_{ESS}, (1-\theta)N_{ESS})$	$Se1, Se2, Sp1, Sp2 \sim Uniform(0,1)$

List of Figures

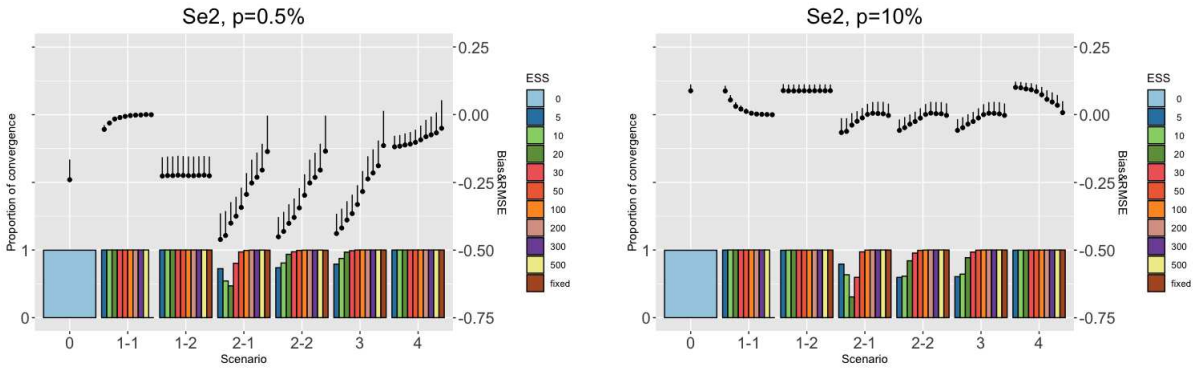

Figure 1: Results at $Se2$ for simulation 1 at prevalence rates of 0.5% (left panel) and 10% (right panel). The black dots and straight lines represent the bias and RMSE of $Se2$, respectively, and the bars represent the convergence rate of the Markov chain Monte Carlo.

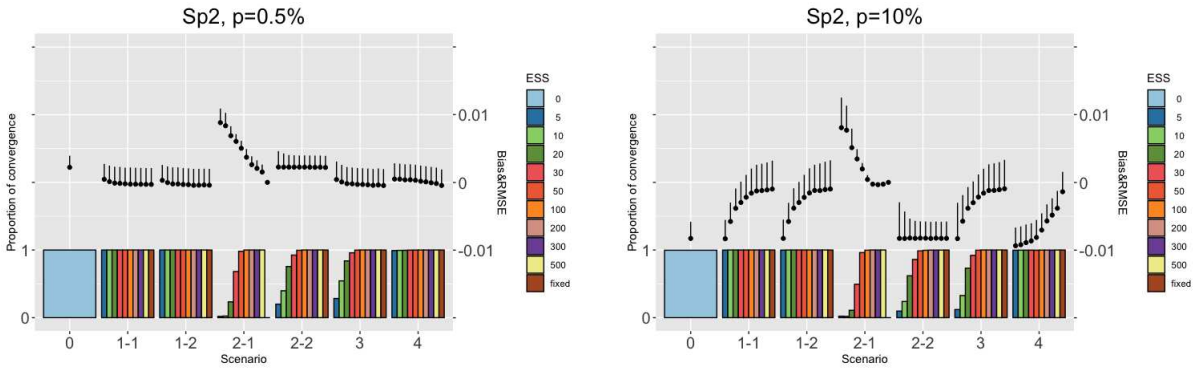

Figure 2: Results at $Sp2$ for simulation 1 at prevalence rates of 0.5% (left panel) and 10% (right panel). The black dots represent the bias of $Sp2$, and the length of the straight line represents the RMSE of $Sp2$. The bars colored by the ESS represent the convergence rate of the Markov chain Monte Carlo.

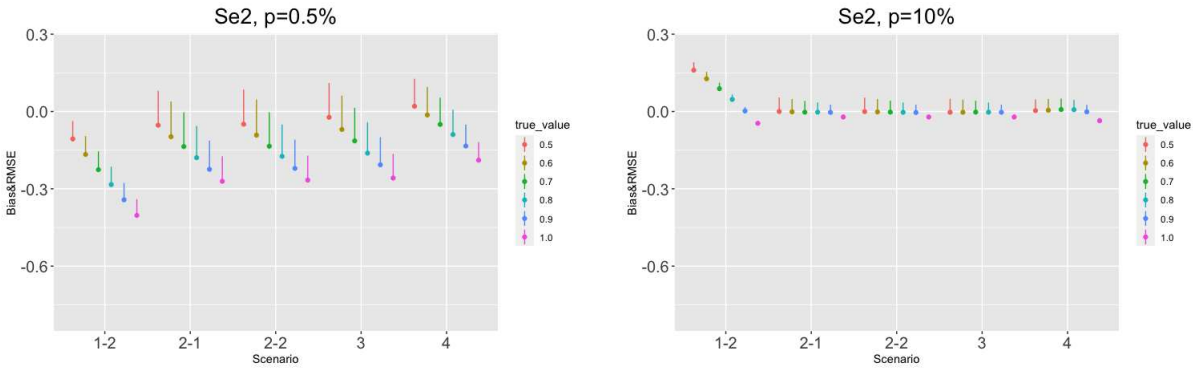

Figure 3: Results at $Se2$ for simulation 2 at prevalence rates of 0.5% (left panel) and 10% (right panel). The dots and straight lines represent the bias and RMSE of $Se2$, respectively.

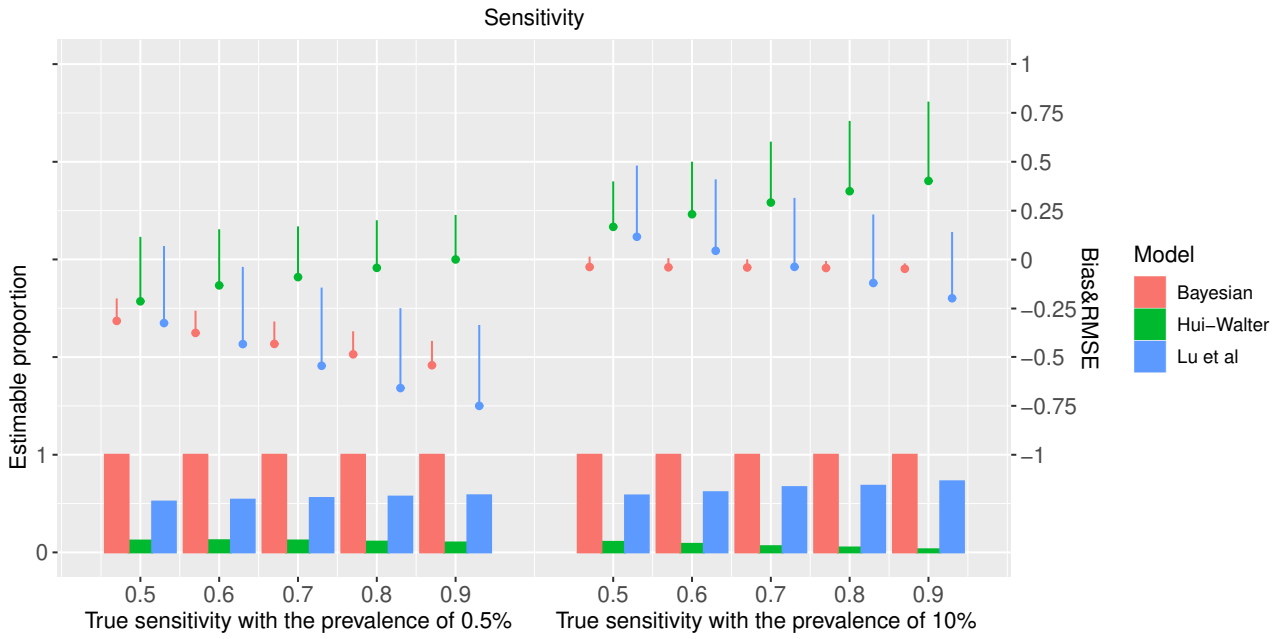

Figure 4: Results at $Se2$ for simulation 3 at prevalence rates of 0.5% (left) and 10% (right). The dots and straight lines represent the bias and RMSE of $Se2$, respectively.