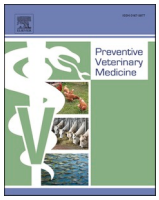
Preventive Veterinary Medicine 217 (2023) 105972

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed



Evaluation of Bayesian Hui-Walter and logistic regression latent class models to estimate diagnostic test characteristics with simulated data

Haifang Ni [a,b,*], Gerrit Koop [a], Irene Klugkist [b], Mirjam Nielen [a]

[a] Department Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University, 3508 TD Utrecht, the Netherlands
[b] Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, 3508 TC Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Bayesian latent class model
Sensitivity
Specificity
Multilevel
Simulation

ABSTRACT

Estimation of the accuracy of diagnostic tests in the absence of a gold standard is an important research subject in epidemiology (Dohoo et al., 2009). One of the most used methods the last few decades is the Bayesian Hui-Walter (HW) latent class model (Hui and Walter, 1980). However, the classic HW models aggregate the observed individual test results to the population level, and as a result, potentially valuable information from the lower level (s) is not fully incorporated. An alternative approach is the Bayesian logistic regression (LR) latent class model that allows inclusion of individual level covariates (McInturff et al., 2004). In this study, we explored both classic HW and individual level LR latent class models using Bayesian methodology within a simulation context where true disease status and true test properties were predefined. Population prevalences and test characteristics that were realistic for paratuberculosis in cattle (Toft et al., 2005) were used for the simulation. Individual animals were generated to be clustered within herds in two regions. Two tests with binary outcomes were simulated with constant test characteristics across the two regions. On top of the prevalence properties and test characteristics, one animal level binary risk factor was added to the data. The main objective was to compare the performance of Bayesian HW and LR approaches in estimating test sensitivity and specificity in simulated datasets with different population characteristics. Results from various settings showed that LR models provided posterior estimates that were closer to the true values. The LR models that incorporated herd level clustering effects provided the most accurate estimates, in terms of being closest to the true values and having smaller estimation intervals. This work illustrates that individual level LR models are in many situations preferable over classic HW models for estimation of test characteristics in the absence of a gold standard.

1. Introduction

The detection of disease is essential for disease control and disease intervention. An ideal situation is to use a perfect diagnostic test with both sensitivity (*Se*) and specificity (*Sp*) of 100%. However for most diseases, there are only imperfect tests available (e.g., Collins and Huynh, 2014; Johnson et al., 2019). In the absence of a perfect (gold standard) reference test, it is challenging to evaluate diagnostic accuracy of the imperfect tests. One of the methods that has often been applied the last few decades is Hui-Walter latent class modelling (Hui and Walter, 1980). This approach links the observed test results from the imperfect diagnostic tests to the unobserved (i.e., latent) disease status. Estimates of the test sensitivity, specificity and disease prevalence can be obtained by using maximum likelihood or Bayesian estimation with one population or more populations with distinct prevalences (Dohoo et al., 2009). One of the limitations of this latent class method is that it aggregates the observed test results from the individual level at the population level. As a result, potentially valuable information from the lower level(s), such as clustering effects within each population and individual level covariates, is not incorporated in the model.

An alternative approach is the logistic regression (LR) latent class model which incorporates the true disease status based on imperfect test results into a LR model (Magder and Hughes, 1997; McInturff et al., 2004). The LR latent class model that allows inclusion of multilevel data can be considered as an extended version of the classic HW model. This approach has been applied under a Bayesian framework in different epidemiologic studies (e.g., McInturff et al., 2004; Lewis et al., 2012; Koop et al., 2013; Hartnack et al., 2013; Paul et al., 2014; O'Hagan et al., 2019; Fernandes et al., 2019) and yields not only estimates for test characteristics but also estimates for the effect of the risk factors. Studies

* Correspondence to: Padualaan 14, 3584 CH Utrecht, the Netherlands.
E-mail address: h.ni@uu.nl (H. Ni).

https://doi.org/10.1016/j.prevetmed.2023.105972

Available online 1 July 2023

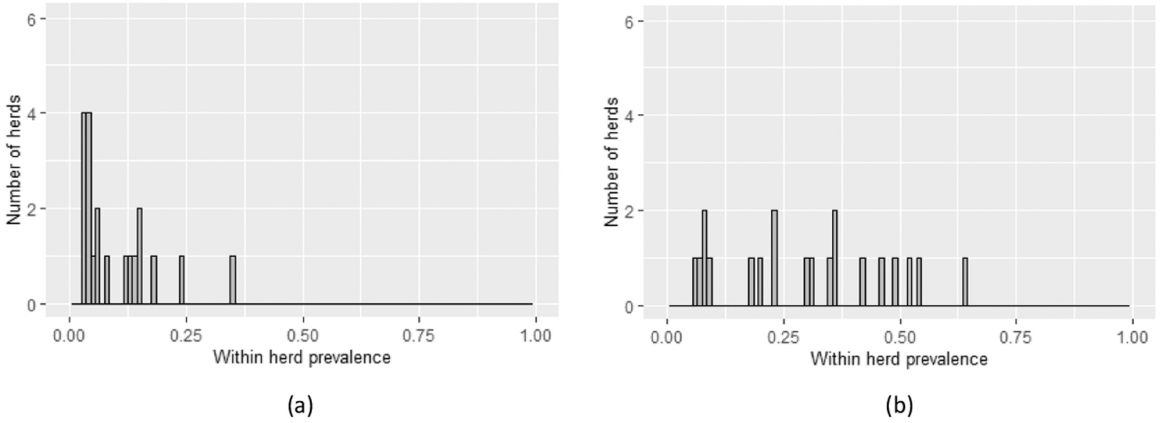


Fig. 1. Within herd prevalences of the 20 herds per region under an intraclass correlation coefficient (ICC) of 0.20 at the herd level: (1a) region 1 with an overall animal prevalence of 10%; (1b) region 2 with an overall animal prevalence of 30%.

that used both classic HW and LR approaches showed that LR models tended to provide more precise posterior estimates for test sensitivity and specificity (Koop et al., 2013; O'Hagan et al., 2019). However in empirical examples, evaluation of these two methods with various settings of population characteristics can be difficult concerning the amount of data collection. Furthermore, assessments of bias and precision of parameter estimates regarding the true value is difficult in real-world research, as the true disease status and true test characteristics are unknown.

In this study, we therefore explored Bayesian classic HW and Bayesian individual level LR latent class models with simulated data where true disease status and true test properties were known. The main objective was to compare the performance of these two approaches in estimating test characteristics. Diverse population settings were simulated where population prevalence, the herd level clustering structure and strength of the risk factor were varied. In addition, we examined the performance of LR models in estimating the association between the risk factor and the disease.

2. Materials and methods

2.1. Data simulation

In order to evaluate model performance in a realistic context, we use the prevalence properties and test characteristics comparable to paratuberculosis in cattle (Toft et al., 2005). Data from two regions were artificially created with an overall animal disease prevalence of 10% for region 1 and 30% for region 2. Both regions contained 20 equal-sized herds. Within each herd, there were 100 cows which resulted in 4000 cattle in total. Two tests with binary outcomes were generated with constant test characteristics across regions and herds. Similar to the study by Toft et al. (2005), the two tests were conditionally independent given the true disease status, with test 1 having a 70% *Se* and a 99% *Sp*, and test 2 a 75% *Se* and a 95% *Sp*.

On top of the prevalence properties and test characteristics, one animal level covariate was added to the data. We chose a binary risk factor generated from the Bernoulli distribution with a success probability of 0.30. The true value of the odds ratio (OR) for the risk factor was set approximately to 1.5 based on the regression coefficient for the risk factor of 0.40. We assumed moderate herd level clustering effects, with an intraclass correlation coefficient (ICC) of 0.20. The random herd effects were sampled from a normal distribution with a herd variance of 0.822 computed from the ICC value by the formula $\sigma_h^2/\ (\sigma_h^2 + \pi^2/3)$where the error variance is fixed and equivalent to $\pi^2/3$ in a logistic regression model (Hox, 2002). The true disease probability for each animal was subsequently calculated using the LR model that included the risk factor with known OR (computed from the known regression coefficient for the risk factor) and the random herd effects. The true binary disease outcome of each animal was then sampled from a Bernoulli distribution with its true disease probability. By adjusting the value of the fixed intercept of the logistic regression, we set the overall animal prevalence for region 1 and region 2 approximately at 10% and 30% respectively. Fig. 1 presents the distributions of within herd prevalences in the two regions for the default data setting.

Table 1
Probability of the 4 test result combinations within one population. The probabilities were formulated under the assumptions of the HW latent class model. Sensitivity of test 1 was denoted Se_1 and specificity Sp_1, likewise sensitivity of test 2 was denoted Se_2 and specificity Sp_2. The overall animal prevalence of the population in region 1 was denoted p_1.

Population 1			
	Test 1		
	Positive		Negative
Test 2	Positive	$Se_1Se_2p_1 + (1-Sp_1)(1-Sp_2)(1-p_1)$	$(1-Se_1)Se_2p_1 + Sp_1(1-Sp_2)(1-p_1)$
	Negative	$Se_1(1-Se_2)p_1 + (1-Sp_1)Sp_2(1-p_1)$	$(1-Se_1)(1-Se_2)p_1 + Sp_1Sp_2(1-p_1)$

2.2. Modelling approach

For the HW approach, crosstabulations based on the combinations of individual animal test results were used as input for the model. Within each population, under the assumptions of conditional independence and constant test properties across populations, the test result combinations of the two tests could be presented in a 2 × 2 contingency table. The stratified populations of the 4000 cattle for the HW models in our study were defined on the basis of the region ID, the herd ID or the binary risk factor. Table 1 presents an example within one of the two populations defined by the region ID. Sensitivity and specificity for test 1 were denoted as Se_1 and Sp_1, and for test 2 as Se_2 and Sp_2. The overall animal prevalence of the population in region 1 was denoted p_1. The probability of each of the four test result combinations was expressed as a function of sensitivity, specificity and prevalence of the population.

An LR mixed model was specified for the multilevel data. When all levels of data were incorporated, i.e., region level, herd level and animal level, the regression model was expressed as follows:

$$logit(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr} + u_r + u_h$$

$$u_r \sim \pi(0, \sigma_r^2)$$

Table 2
Ten Bayesian latent class models to estimate the sensitivity and specificity of two imperfect tests. The region level is subscripted as *r*, the herd level as *h* and RF represents the risk factor.

	Model specification
HW model	
HW_r	two equal-sized populations defined by region ID
HW_h	40 equal-sized populations defined by herd ID
HW_RF	Two unequal-sized populations defined by the binary RF
LR model	
LR_r	$logit(p_{ihr}) = \beta_0 + u_r$
LR_h	$logit(p_{ihr}) = \beta_0 + u_h$
LR_RF	$logit(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr}$
LR_r_h	$logit(p_{ihr}) = \beta_0 + u_r + u_h$
LR_RF_r	$logit(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr} + u_r$
LR_RF_h	$logit(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr} + u_h$
LR_RF_r_h	$logit(p_{ihr}) = \beta_0 + \beta_1 RF_{ihr} + u_r + u_h$

$u_h \sim \pi(0, \sigma_h^2)$.

The risk factor at individual level was denoted RF_{ihr} and the latent underlying disease probability for the observed binary outcome was denoted p_{ihr} for individual $i(i = 1, \ldots, n_{hr})$in herd $h(h = 1, \ldots, H)$ from region $r(r = 1, 2)$. The random region effects and the random herd effects were assumed to have a normal distribution with mean zero and variance σ_r^2 for the regions and variance σ_h^2 for the herds. The disease probability of each animal p_{ihr} was estimated by the LR mixed model. Instead of population level crosstabulations as in HW models, with LR models, a crosstabulation was constructed at the individual level. In our example with two imperfect diagnostic tests, probabilities of the four test result combinations as shown in Table 1 could be expressed with the (latent) disease probability of each animal p_{ihr}, the sensitivity and specificity of the two tests.

2.3. Analysis of simulated data

Ten latent class models were specified under the Bayesian framework for estimation of the test characteristics. Table 2 presents the specification for these models. Three HW models were applied, stratifying on region ID (*HW_r*), herd ID (*HW_h*) and the binary risk factor (*HW_RF*). The crosstabulations for the test result combinations can be found in Table A1 of the Supplementary information. These crosstabulations were used as input data for the corresponding HW models. Seven LR models were specified with one or more levels of data (i.e., individual, herd, and population level) incorporated. Comparisons were made between the HW models and their corresponding LR models (e.g., *HW_r* and *LR_r*) as well as between the seven LR models.

For parameter estimation, non-informative beta prior distributions $beta(1, 1)$ were assigned to all four sensitivity and specificity parameters and the region prevalences in both modelling approaches. For the LR models, non-informative normal prior distributions with mean 0 and a large variance $N(0, 1000)$ were specified for the regression coefficients (β_0, β_1), and non-informative inverse-gamma prior distributions $inverse\text{-}gamma(0.001, 0.001)$ were specified for the variance of the random region effects (u_r) and random herd effects (u_h). Four Markov chain Monte Carlo (MCMC) posterior chains were sampled for each model using JAGS (Plummer, 2003) called from R (R Core Team, 2016) by using the 'runjags' package (Denwood, 2016). Within each chain, the first 5000 iterations were discarded as the burn-in phase and the subsequent 10,000 iterations were saved for parameter inferences. Convergence was checked using psrf values and traceplots. Posterior distributions were only used after ensuring that convergence was reached.

2.4. Sensitivity analysis with varying population characteristics

Several sensitivity analyses were performed by varying the population characteristics of the simulated datasets. In order to investigate the impact of the risk factor, animals from the default data setting were permuted between the two categories of the risk factor within each region to model a higher OR (i.e., 2.7, 7.4). To examine the effect of herd level clustering, animals from the default data setting were permuted among the herds within each region resulting in a lower (0.10) or a higher ICC (0.30). Further investigations on the size of the region prevalences and the difference between the region prevalences were not done by permuting the original dataset, but based on new simulated datasets, as the overall animal prevalence changed in these settings in comparison to the default setting. Impact of the region prevalences was evaluated in a lower prevalence range (5%, 25%) and in a higher range (30%,50%) while keeping the difference between the region prevalences 20% as in the default setting. The effect of the difference between region prevalences was evaluated as well by changing the difference to 10% (10%, 20%) and 40% (10%, 50%).

2.5. Model comparisons

Within each data setting, posterior estimates for the test sensitivity and specificity of the two tests were obtained for the ten Bayesian latent class models presented in Table 2. As in veterinary epidemiology, posterior results are often summarized by means of posterior median and the 95% credible interval, in our study, we adopted the same approach for presenting our results. The difference between the posterior median and the true parameter value is referred to as the bias of the estimate and the width of the 95% posterior credible interval is referred to as the precision. Note that the terms bias and precision are thus used while evaluating just one synthetic dataset. Investigating different settings with one simulated dataset per setting is not uncommon (see, for instance, Mulder et al., 2009) and, in the context of our study, serves the purpose of making mutual comparisons between estimation methods on the same data. As such we are interested in the relative performance of different potential models for the estimation and not so much in the absolute performance. Furthermore, posterior estimates for the regression coefficient of the risk factor were evaluated in all data settings for the four LR models that included the risk factor.

3. Results and discussion

The HW and LR modelling approaches were first examined in the default setting, followed by sensitivity analyses with varying population characteristics. Posterior estimates for test sensitivity and specificity of the two tests are graphically summarized in separate plots. The vertical lines in the plots represent the true values of the four test properties. The horizontal intervals represent the 95% posterior credible intervals, and the squares within the interval represent the posterior medians. The first six models (above the dashed line) incorporated either only region, herd or the risk factor using the HW and LR approach. The last four LR models included two or all three levels of data.

For the sensitivity analyses, in order to compare between various data settings, results from the default setting are added to the plots, with the grey bars displaying the 95% credible intervals and the grey squares representing the medians. Numeric summaries for the results are available in Tables A2-A7 of the Supplementary information.

Please note, again, that we use the terms bias and precision to refer to estimation results from a single dataset. Therefore, the results cannot be interpreted as the absolute bias and variance of the estimators. Instead, the results provide a measure for how close the estimates are to the true values and how precise the estimates are in terms of the precision (range) of the 95% posterior credible intervals and inform us about the relative performance when comparing the different modeling approaches.

In this simulation study, we compared the two methods under data settings with different population characteristics and we modelled conditionally independent diagnostic tests. This is a limitation of our

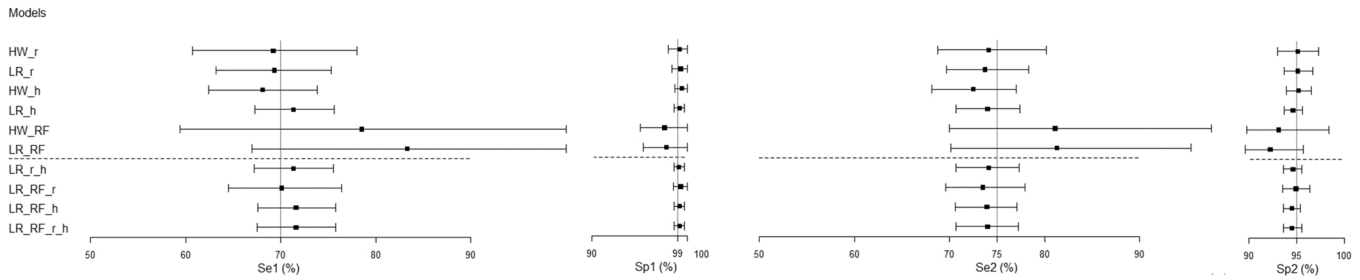


Fig. 2. Summary plots for posterior estimates of the test characteristics under the default data setting. Two regions contain in total 4000 cattle, with each region consisting of 20 equal sized herds and each herd consisting of 100 cows. The overall animal prevalences for the two regions are 10% and 30% respectively, and the intraclass correlation coefficient (ICC) at the herd level is 0.20. A binary animal level risk factor is present with success probability 0.30 and is associated to the true disease status with an odds ratio (OR) 1.5. True test characteristics are represented by the grey vertical lines. Ten Bayesian latent class models are specified, with three Hui-Walter (HW) models stratifying on region ID (*r*), herd ID (*h*) and the risk factor (*RF*) and seven logistic regression (LR) models incorporating data on one or more levels (i.e., region, herd, animal level).

study and future studies may explore these questions in the context of conditionally dependent diagnostic tests.

Model convergence checks showed that for the default setting all psrf were below 1.1. However, for some parameters in the other settings this was not the case. We rerun these models with more iterations (for each chain, 10,000 burn-in followed by 50,000 for inference) to obtain acceptable psrf values and also monitored the traceplots of the suspect parameters to ensure that all reported estimates were reliable. All inspected traceplots showed proper convergence and the parameter estimates did not substantially change with increased iterations.

3.1. Default data setting

Fig. 2 presents the posterior estimates for the default setting. This figure clearly shows that all ten models provided less biased and more precise estimates for test specificities than for test sensitivities. This was in line with the fact that there was less data available to estimate sensitivity than there was to estimate specificity as the overall animal prevalences were lower than 50% within the two regions (10%,30%). In addition, the test specificities were much higher ($Sp_1 = 99\%$, $Sp_2 = 95\%$) than the test sensitivities ($Se_1 = 70\%, Se_2 = 75\%$).

Further, for all estimates from the ten models, the true values were located within the 95% credible intervals. One can see that for the first six models, which only included region, herd or risk factor, the credible intervals for the LR models were narrower than the corresponding HW models, with the LR model that incorporated only herd level clustering effects (*LR_h*) showing the best precision. It is notable that the HW and LR models that only incorporated data on the risk factor produced wide credible intervals. The strikingly poorer performance of these two models relative to the other eight models led to further investigation on the risk factor regarding the sample sizes and prevalence characteristics of the populations defined by the risk factor (see 3.6).

The last four LR models that included two or all three levels showed smaller credible intervals in comparison to the first six models except for the *LR_h* model. Posterior estimates from LR models that incorporated herd level clustering effects (i.e., *LR_h, LR_r_h, LR_RF_h, LR_RF_r_h*) all

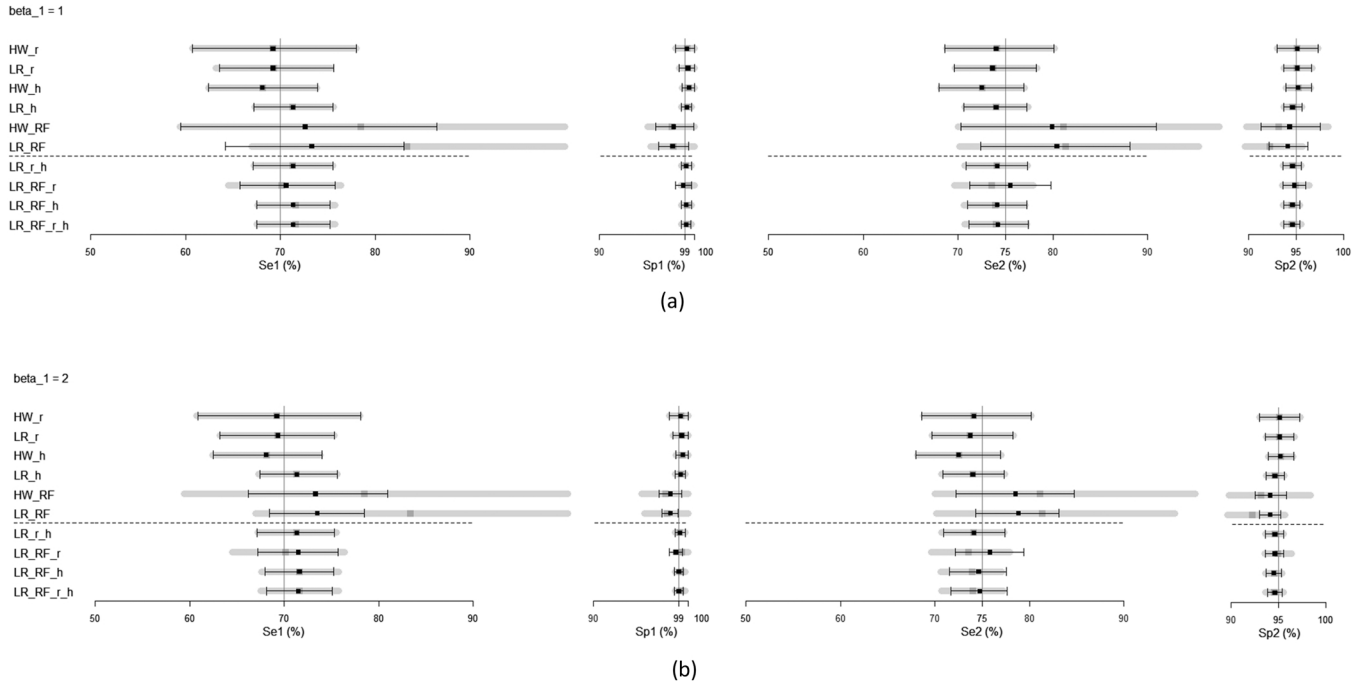


Fig. 3. Summary plots for posterior estimates of the evaluation of effect of strength of the association between the animal level risk factor and the disease on test sensitivity and specificity estimation. The animal level risk factor is binary and has success probability 0.30. The upper panel (3a) presents odds ratio (OR) = 2.7 ($\beta_1 = 1$) and the bottom panel (3b) presents OR = 7.4 ($\beta_1 = 2$). The grey squares and bars represent results from the default data setting OR = 1.5 ($\beta_1 = 0.4$). See Fig. 2 for further details of the population characteristics.

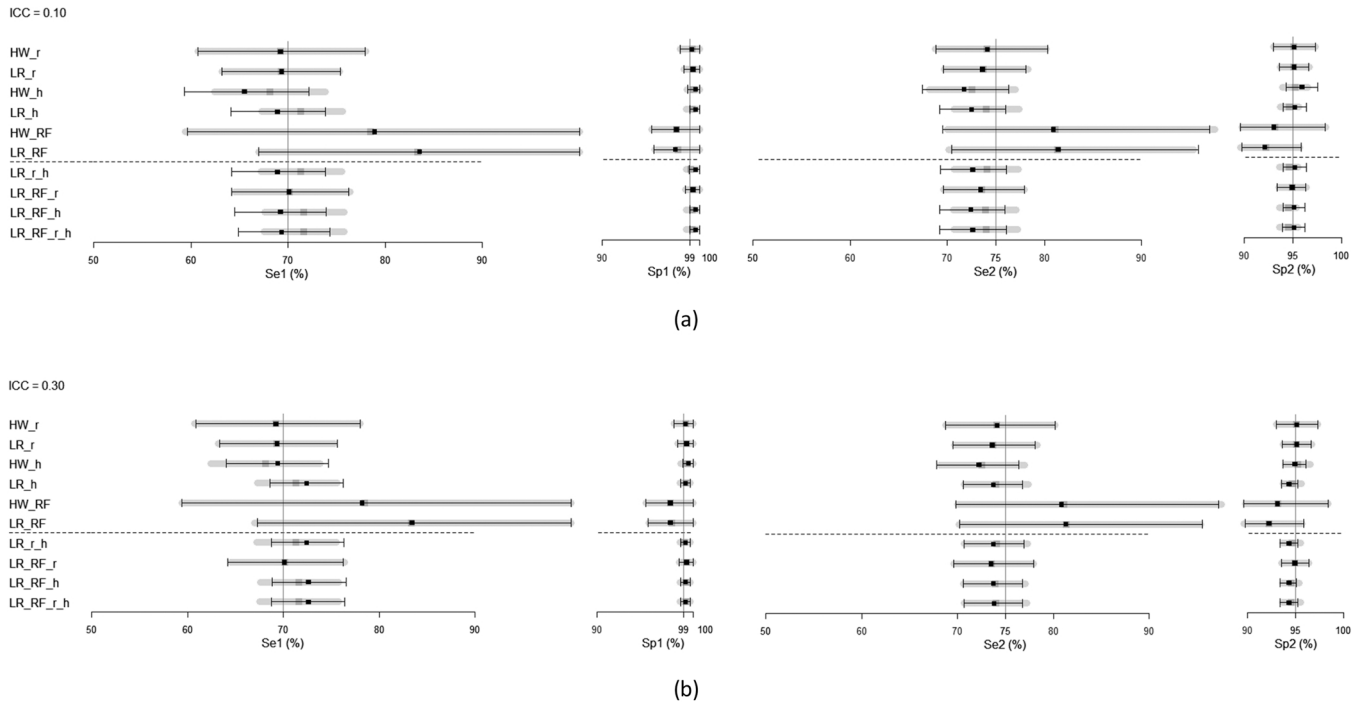


Fig. 4. Summary plots for posterior estimates of the evaluation of effect of strength of intraclass correlation coefficient (ICC) at the herd level on test sensitivity and specificity estimation. The upper panel (4a) presents ICC = 0.10 and the bottom panel (4b) presents ICC = 0.30. The grey squares and bars represent results from the default data setting ICC = 0.20. See Fig. 2 for further details of the population characteristics.

showed similar bias and precision.

The difference in precision from the HW and LR models was also observed in other studies that applied both methods. In the study by Koop et al. (2013) for instance, when evaluating the test performance of bacteriological culture and somatic cell counts for subclinical intramammary infection in Dutch goats, authors observed narrower posterior credible intervals from the LR models in comparison to the HW model. In addition, the LR model that included most risk factors (i.e., 3) provided the narrowest credible intervals. Likewise O'Hagan et al. (2019) also reported that the LR model with risk factors showed narrower credible intervals than the HW model for sensitivity and specificity estimates of the single intradermal comparative cervical tuberculin test and post-mortem examination for bovine tuberculosis in cattle from Northern Ireland.

3.2. *Effect of different associations between the risk factor and disease*

We further investigated the impact of a stronger association between the risk factor and disease on the estimates of test characteristics. Bias and precision of the posterior estimates from each model are presented in Fig. 3. For models that only incorporated data on region level or herd level, results remained the same as those from the default dataset. This was expected, as for this sensitivity analysis animals simulated under the default setting were permuted between the two categories of the risk factor but remained in the same regions and herds.

For the HW model that defined populations on the basis of the risk factor, estimates were less biased and more precise when the regression coefficient for the risk factor increased from 0.40 (default) to 1, and from 1 to 2. A possible explanation for this finding was that populations stratified by the risk factor had more distinct population prevalences, when the risk factor had a stronger association with the disease status. Similar improvement was seen in the LR model that only included the risk factor. This may be because more variance of the data was explained at the animal level by the risk factor when the regression coefficient is stronger.

Defining populations on the basis of risk factors should be done with caution for HW models. Results of this sensitivity analysis indicated the necessity of checking the strength of the association between the risk factor and the disease status when using the HW approach. Posterior estimates of the HW model were less biased and more precise when the individual level risk factor had a stronger association with the disease. In veterinary epidemiology, individual level risk factors such as history of mastitis in previous lactations for bovine mastitis (Jamali et al., 2018) and body condition score for ketosis in cows (Vanholder et al., 2015), herd level risk factors such as direct cattle importation for paratuberculosis (Rangel et al., 2015) and herd size for bovine tuberculosis (Bessell et al., 2012) are found to have relatively strong association with the respective diseases within the target populations (ORs ranging from 2.06 to 19.22). However, Toft et al. (2005) pointed out that defining populations based on individual level biological risk factors such as age may violate the HW model assumption of constant test characteristics across the stratified populations due to for instance cross reactions. Higher level geographic risk factors such as zip-code and veterinary practices that result in populations with distinct prevalences are often preferred as stratifiers. Based on results of this sensitivity analysis, we recommend researchers to choose the LR approach when risk factors are available.

3.3. *Effect of strength of herd level clustering (ICC)*

In Fig. 4, results from datasets with varying strength of herd level clustering effects are presented. Animals simulated under the default setting were permuted between herds but remained in the same regions and risk factor categories. Therefore, the models that did not incorporate herd level clustering effects produced the same results as the default dataset. For models that incorporated herd level effects, when the ICC was reduced from the default 0.20 to 0.10 (4a), posterior estimates were slightly more biased and less precise. However, LR models that incorporated herd level effects as well as the risk factor and/or the region effects showed still reasonable estimates. When the ICC increased from the default 0.20 to 0.30 (4b), the bias of the posterior estimates was similar to the default setting but the precision increased slightly.

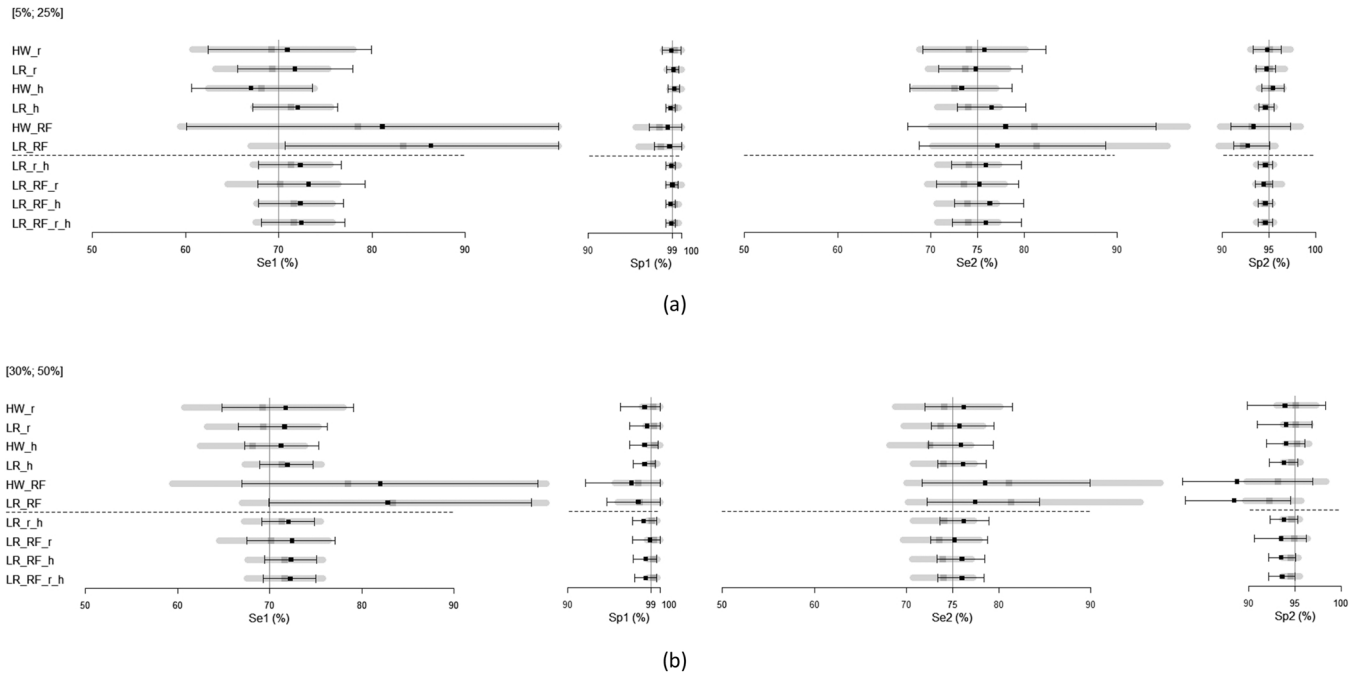


Fig. 5. Summary plots for posterior estimates of the evaluation of the size of region prevalences on test sensitivity and specificity estimation. The upper panel (5a) presents lower region prevalences (5%, 25%) and the bottom panel (5b) presents higher region prevalences (30%, 50%). The grey squares and bars represent results from the default data setting (10%, 30%). See Fig. 2 for further details of the population characteristics.

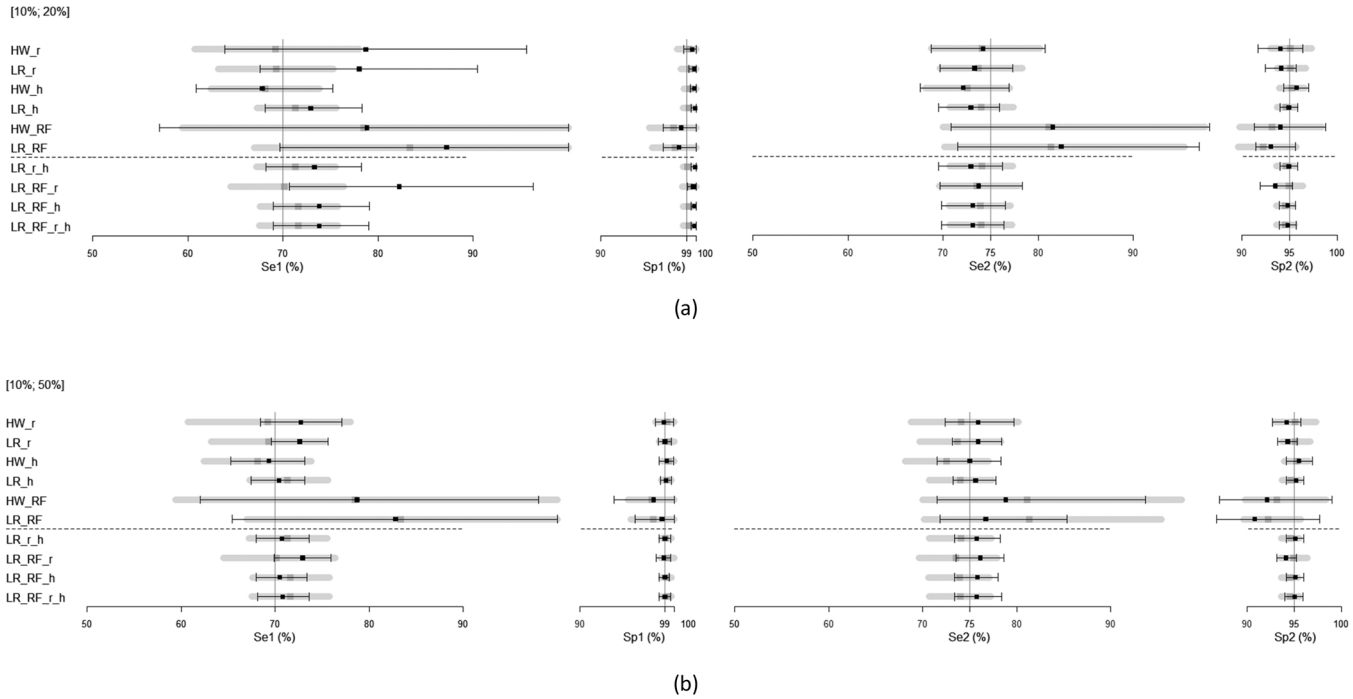


Fig. 6. Summary plots for posterior estimates of the evaluation of the difference in region prevalences on test sensitivity and specificity estimation. The upper panel (6a) presents 10% difference (10%, 20%) and the bottom panel (6b) presents 40% difference (10%, 50%). The grey squares and bars represent results from the default data setting (10%, 30%). See Fig. 2 for further details of the population characteristics.

In veterinary epidemiology, herd level clustering effects have been computed for various infectious diseases and ICC values are found to vary from 0.04 (Anaplasma marginale in cattle) to 0.42 (bovine viral diarrhea in cattle), and most diseases have an ICC below 0.20 (Dohoo et al., 2009). Results in this sensitivity analysis suggest that it might still be useful to include herd level clustering effects in the latent class models for the estimation of diagnostic test characteristics even when the ICC value is relatively low.

3.4. Effect of different values for region prevalences

Fig. 5 summarizes the effect of varying the overall animal prevalences of the two regions while keeping the difference constant. For this analysis, two new datasets were generated as the region prevalences

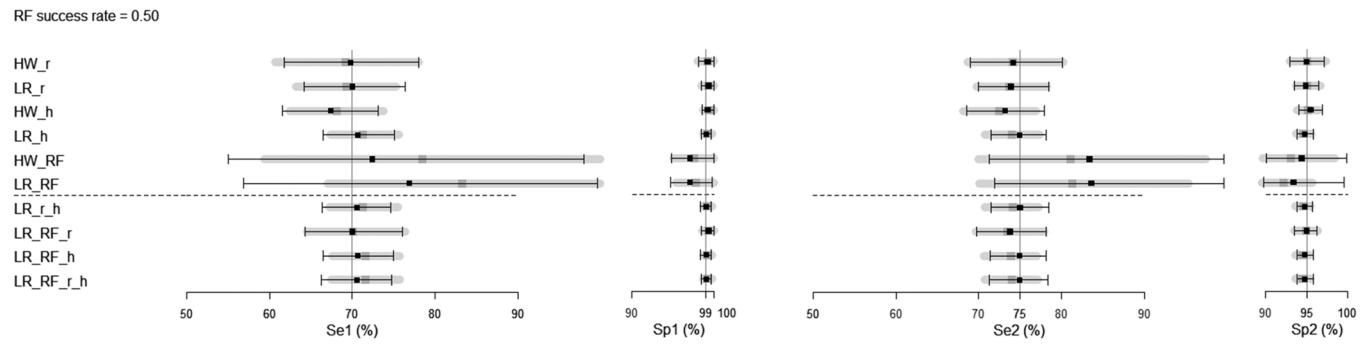


Fig. 7. Summary plots for posterior estimates of the test characteristics with the binary risk factor sampled from success probability 0.50 with an odds ratio of 1.5. Two regions contain in total 4000 cattle, with each region consisting of 20 equal sized herds and each herd consisting of 100 cows. The grey squares and bars represent results from the default data setting with success probability 0.30 for the risk factor. See Fig. 2 for further details of the population characteristics.

were changed in comparison to the default dataset. When the region prevalences were reduced from the default 10% and 30% to 5% and 25% (5a), the precision for test sensitivities worsened whereas the precision for test specificities improved. In contrast, when the animal prevalences of the regions were increased from the default setting to 30% and 50% (5b), the precision for test sensitivities improved and for test specificities worsened. This is expected, because when the animal prevalence is lower, there is less information available in the data to estimate test sensitivity. Likewise when the animal prevalence is higher, the amount of information for estimation of test sensitivity increased.

The performance of the models regarding bias and precision from the dataset with region prevalences 5% and 25% was comparable to the default setting. This indicates that HW and LR approaches are robust if one of the populations has a low prevalence, as long as the difference between the population prevalences is distinct. The LR models that incorporated herd level clustering effects showed again the least biased and the most precise posterior estimates in comparison to other HW and LR models.

3.5. Effect of the difference between region prevalences

Fig. 6 contains results of the ten models for data settings where difference between the animal prevalences of the two regions was changed from 20% to 10% or 40%. These results were also based on two new datasets as the region prevalences were changed from the default dataset. When the region prevalences were changed from the default (10%, 30%) to (10%, 20%) (6a), precision of the estimates for test sensitivities from all models worsened, whereas precision of the estimates for test specificities improved. This was due to less data available to estimate sensitivity than to estimate specificity as the overall animal prevalence was smaller than in the default setting. However, when the region prevalences were changed from the default (10%, 30%) to (10%, 50%) (6b), precision of the estimates for test sensitivities from all models improved, whereas precision of the estimates for test specificities worsened.

It is unclear based on the results of this sensitivity analysis, whether the change in model performance that included region information (*HW_r*, *LR_r*, *LR_r_h*, *LR_RF_r*, *LR_RF_r_h*) was fully due to the change in the prevalence difference between the two regions. In the next section we further investigated the effect of different population prevalences and sample sizes on model performance.

3.6. Difference in population prevalences and sizes based on the risk factor

In the sensitivity analyses we presented above, it is clear that HW and LR models with only the risk factor showed the largest bias and the worst precision (Figs. 2–6). In order to grasp whether results from these two models were influenced by unequal population sizes, we simulated one more dataset where the success probability of the binary risk factor was

Table 3
The prevalences and sample sizes of the two stratified populations based on the individual level risk factor within each data setting.

	Population prevalence (population size)		Difference between prevalences	Figure number
	RF = 0	RF = 1		
Default	18.0% (2760)	24.2% (1240)	6.2%	Fig. 2
$\beta_1 = 1$	15.6% (2760)	29.4% (1240)	13.8%	Fig. 3a
$\beta_1 = 2$	11.1% (2760)	38.8% (1240)	27.7%	Fig. 3b
ICC = 0.10	18.0% (2760)	24.2% (1240)	6.2%	Fig. 4a
ICC = 0.30	18.0% (2760)	24.2% (1240)	6.2%	Fig. 4b
Region prevalences = (5%, 25%)	13.7% (2798)	18.5% (1202)	4.8%	Fig. 5a
Region prevalences = (30%, 50%)	37.5% (2826)	46.0% (1174)	8.5%	Fig. 5b
Region prevalences = (10%, 20%)	13.9% (2813)	17.9% (1187)	4.0%	Fig. 6a
Region prevalences = (10%, 50%)	28.3% (2761)	34.1% (1239)	5.8%	Fig. 6b
RF success probability = 0.50	17.2% (1933)	22.5% (2067)	5.3%	Fig. 7

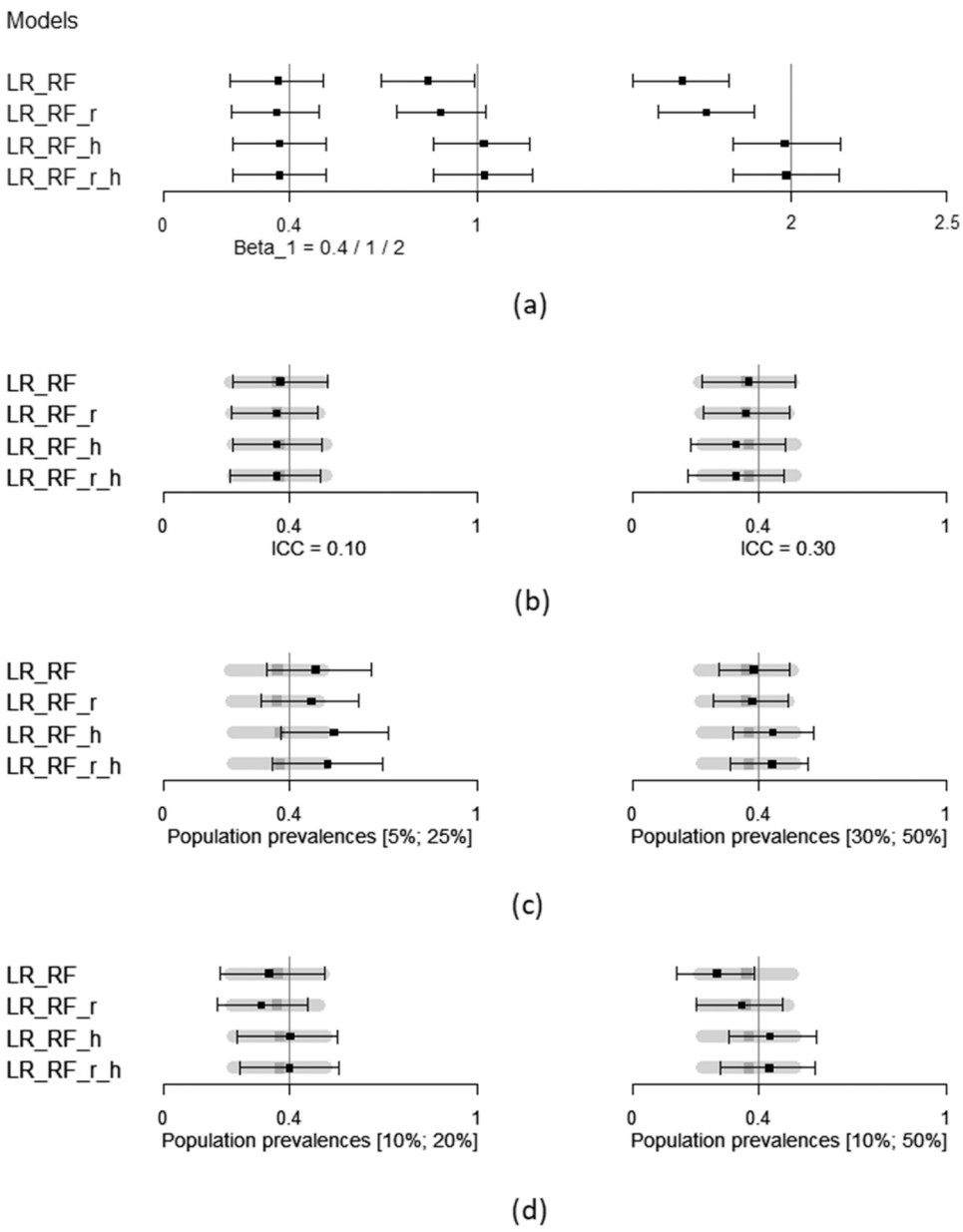


Fig. 8. Summary plots for posterior estimates of the regression coefficient β_1 of the risk factor in different data settings from four Bayesian logistic regression latent class models that either include risk factor only, or include risk factor and region and/or herd level effects. The binary risk factor at the animal level has success probability 0.30. The first panel (8a) presents estimates from data settings that have true β_1 values of 0.4 (default setting), 1 and 2, corresponding to odd ratios (OR) of 1.4, 2.7 and 7.4 respectively. The following panels 8b to 8d present results from data settings that have true β_1 value of 0.4 but with either different intra-class correlation coefficient (ICC) at the herd level or different region prevalences. The grey squares and bars represent results from the default setting.

changed from 0.30 (default) to 0.50. With success probability 0.50, the difference between the sample sizes of the two risk factor categories was minimal.

Estimates from the *HW_RF* and *LR_RF* models became slightly better regarding bias but worse regarding precision compared to the default setting (Fig. 7). These results verified that unbalanced sample sizes of the stratified populations were not the main cause of the relatively poor precision of posterior estimates from the models. In order to further understand the impact of population prevalences and population sample sizes on posterior estimates for test characteristics, we listed the values of prevalences and sample sizes of populations stratified by the risk factor from all data settings. Table 3 shows that in the setting with the risk factor sampled from the success probability 0.50, the sample sizes were indeed similar, however the population prevalence difference was reduced from the default 6.2–5.3%. In fact, for most data settings, when we split the data on the basis of the risk factor, population prevalence differences were below 10%, with the exception of the two settings where the regression coefficient for the individual level risk factor was relatively strong (corresponding to an OR of 2.7 and 7.4 respectively). The effect of a small population prevalence difference on large posterior estimate credible intervals was also reported in Johnson et al. (2019).

However, it is still possible that the unbalanced sample sizes of the stratified populations also played a role in the poor performance of the HW and LR models that only used the risk factor information. Future studies should further investigate the effect of unbalanced population sample sizes on HW and LR modelling approaches.

3.7. Estimates of the regression coefficient

In order to obtain the estimate of the association between the individual level risk factor and disease, we examined the posterior results of the regression coefficient from the LR models that included the risk factor. In Fig. 8, one can see at the upper panel (8a), the true regression coefficient value was changed from 0.40 to 1 and 2, with 0.40 presenting the default setting. Results showed that the LR model without herd level clustering effects incorporated (*LR_RF*, *LR_RF_r*) tended to underestimate the association between the risk factor and the disease. This phenomenon has been shown before and was explained by Hedeker and Gibbons,

2006). Estimates for the regression coefficients of covariates from a fixed-effects LR model tend to be closer to zero than those resulted from a mixed-effects (i.e., random effects) LR model. Estimates from the fixed-effects are considered "population-averaged" which indicate the effect of covariates averaging over the population (in our example, over the herds), whereas estimates from the mixed-effects LR model are "cluster-specific" since they are conditional on the random clustering effects. Results in this study showed that in datasets with moderate clustering effects (ICC = 0.20), when the association between the risk factor and the disease status was weak, population averaged and cluster specific posterior estimates for the regression coefficient of the risk factor were similar regarding bias and precision. However, when the association was stronger, the cluster specific estimates were much less biased in comparison to the population averaged estimates.

The panels 8b to 8d showed six data settings with other population characteristics, but always with a default regression coefficient of 0.40. Performance of the four LR models on estimation of the association between the risk factor and the disease was similar across various settings, with the exception of the one with low overall animal prevalence (i.e., 5%, 25%). The deviation of the estimates in this setting might be caused by a lack of information on the association between the risk factor and the disease status from region 1 as the prevalence was only 5%.

4. Conclusion

HW and LR latent class models are two approaches to estimate test characteristics when the true disease status is unknown and when there is no gold standard. We show that LR models applied in a setting of two conditionally independent tests are more precise in posterior estimates across various settings, with the LR models that incorporated herd level clustering effects presenting the least biased and most precise estimates. Results also revealed that stratifying data on the basis of an individual level risk factor for the HW modelling approach can be problematic, unless one is certain that the association between the risk factor and the outcome is strong. Altogether, this work shows that LR models are in many situations the preferable alternative to HW models to estimate test characteristics in the absence of a perfect reference test.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.prevetmed.2023.105972.

References

Bessell, P.R., Orton, R., White, P.C.L., Hutchings, M.R., Kao, R.R., 2012. Risk factors for bovine Tuberculosis at the national level in Great Britain. BMC Vet. Res 8, 51. https://doi.org/10.1186/1746-6148-8-51.

Collins, J., Huynh, M., 2014. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. Stat. Med. 33, 4141–4169. https://doi.org/10.1002/sim.6218.

Denwood, M.J., 2016. Runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. J. Stat. Softw. 71. https://doi.org/10.18637/jss.v071.i09.

Dohoo, I.R., Martin, W., Stryhn, H.E., 2009. Veterinary Epidemiologic Research, 2nd ed.,. VRC Inc.

Fernandes, L.G., Denwood, M.J., Santos, C.S.A.B., Alves, C.J., Pituco, E.M., Romaldini, A. H.C.N., De Stefano, E., Nielsen, S.S., De Azevedo, S.S., 2019. Bayesian estimation of herd-level prevalence and risk factors associated with BoHV-1 infection in cattle herds in the State of Paraíba, Brazil. Prev. Vet. Med 169, 104705. https://doi.org/10.1016/j.prevetmed.2019.104705.

Hartnack, S., Budke, C.M., Craig, P.S., Jiamin, Q., Boufana, B., Campos-Ponce, M., Torgerson, P.R., 2013. Latent-class methods to evaluate diagnostics tests for Echinococcus infections in dogs. PLoS Negl. Trop. Dis. 7, e2068 https://doi.org/10.1371/journal.pntd.0002068.

Hedeker, D., Gibbons, R.D., 2006. Longitudinal Data Analysis. John Wiley & Sons, New Jersey.

Hox, J.J., 2002. Multilevel Analysis: Techniques and Applications. Lawrence Erlbaum Associations, New Jersey.

Hui, S., Walter, S., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171. https://doi.org/10.2307/2530508.

Jamali, H., Barkema, H.W., Jacques, M., Lavallée-Bourget, E.M., Malouin, F., Saini, V., Stryhn, H., Dufour, S., 2018. Invited review: Incidence, risk factors, and effects of clinical mastitis recurrence in dairy cows. J. Dairy Sci. 101, 4729–4746. https://doi.org/10.3168/jds.2017-13730.

Johnson, W.O., Jones, G., Gardner, I.A., 2019. Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies. Prev. Vet. Med. 167, 113–127. https://doi.org/10.1016/j.prevetmed.2019.01.010.

Koop, G., Collar, C.A., Toft, N., Nielen, M., Van Werven, T., Bacon, D.A.C., Gardner, I.A., 2013. Risk factors for subclinical intramammary infection in dairy goats in two longitudinal field studies evaluated by Bayesian logistic regression. Prev. Vet. Med 108, 304–312. https://doi.org/10.1016/j.prevetmed.2012.11.007.

Lewis, F., Sanchez-Vazquez, M.J., Torgerson, P.R., 2012. Association between covariates and disease occurrence in the presence of diagnostic error. Epidemiol. Infect. 140 (8), 1515–1524 (Aug).

Magder, L.S., Hughes, J.P., 1997. Logistic regression when the outcome is measured with uncertainty. Am. J. Epidemiol. 146, 195–203. https://doi.org/10.1093/oxfordjournals.aje.a009251.

McInturff, P., Johnson, W.O., Cowling, D., Gardner, I.A., 2004. Modelling risk when binary outcomes are subject to error. Stat. Med. 23, 1095–1109. https://doi.org/10.1002/sim.1656.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W.H.J., Selfhout, M., Hoijtink, H., 2009. Bayesian model selection of informative hypotheses for repeated measurements. J. Math. Psychol. 53, 530–546. https://doi.org/10.1016/j.jmp.2009.09.003.

O'Hagan, M.J.H., Ni, H., Menzies, F.D., Pascual-Linaza, A.V., Georgaki, A., Stegeman, J. A., 2019. Test Characteristics of the tuberculin skin test and post-mortem examination for bovine tuberculosis diagnosis in Cattle in Northern Ireland estimated by bayesian latent class analysis with adjustments for covariates. Epidemiol. Infect. 147, e209 https://doi.org/10.1017/S0950268819000888.

Paul, S., Agger, J.F., Agerholm, J.S., Markussen, B., 2014. Prevalence and risk factors of Coxiella burnetii seropositivity in Danish beef and dairy cattle at slaughter adjusted for test uncertainty. Prev. Med. Vet. 113, 504–511. ⟨https://doi.org/10.1016/j.prevetmed.2014.01.018⟩.

Plummer, M., 2003. JAGS: a program for analysis of bayesian graphical models using gibbs sampling JAGS: just another gibbs sampler. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Pp. March 20–22 ISSN 1609–395X.

R Core Team, 2016. R: a Language and Environment for Statistical Computing. URL. R Foundation for Statistical Computing, Vienna, Austria. ⟨https://www.R-project.org/⟩.

Rangel, S.J., Paré, J., Doré, E., Arango, J.C., Côté, G., Buczinski, S., Labrecque, O., Fairbrother, J.H., Roy, J.P., Wellemans, V., Fecteau, G., 2015. A systematic review of risk factors associated with the introduction of Mycobacterium avium spp. paratuberculosis (MAP) into dairy herds. Can. Vet. J. 56, 169–177.

Toft, N., Jørgensen, E., Højsgaard, S., 2005. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. Prev. Vet. Med. 68, 19–33. https://doi.org/10.1016/j.prevetmed.2005.01.006.

Vanholder, T., Papen, J., Bemers, R., Vertenten, G., Berge, A.C., 2015. Risk factors for subclinical and clinical ketosis and association with production parameters in dairy cows in the Netherlands. J. Dairy Sci. 98, 880–888. https://doi.org/10.3168/jds.2014-8362.