**Title**

# Low accuracy of Bayesian latent class analysis for estimation of herd-level true prevalence under certain disease characteristics - An analysis using simulated data

**Author names and affiliations**

Conor G. McAloon[a], Michael L. Doherty[a], Paul Whyte[a], Cristobal Verdugo[b], Nils Toft[c], Simon J. More[a], Luke O'Grady[a], Martin J. Green[d]

[a]School of Veterinary Medicine, University College Dublin, Belfield, Dublin D04 W6F6, Ireland

[b]Instituto de Medicina Preventiva Veterinaria, Universidad Austral de Chile, Valdivia, Chile

[c]National Veterinary Institute, Technical University of Denmark, Lyngby, Denmark

[d]School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington, United Kingdom

**Corresponding author**

Conor G. McAloon

E-mail: conor.mcaloon@ucd.ie

Section of Herd Health and Animal Husbandry, School of Veterinary Medicine, University College Dublin, Belfield, Dublin D04 W6F6, Ireland

+353 1 716 6083

**Abstract**

Estimation of the true prevalence of infected individuals involves the application of a diagnostic test to a population and adjusting according to test performance, sensitivity and specificity. Bayesian latent class analysis for the estimation of herd and animal-level true prevalence, has become increasingly used in veterinary epidemiology and is particularly useful in incorporating uncertainty and variability into analyses in a flexible framework. However, the approach has not yet been evaluated using simulated data where the true prevalence is known. Furthermore, using this approach, the within-herd true prevalence is often assumed to follow a beta distribution, the parameters of which may be modelled using hyperpriors to incorporate both uncertainty and variability associated with this parameter. Recently however, the authors of the current study highlighted a potential issue with this approach, in particular, with fitting the distributions and a tendency for the resulting distribution to invert and become clustered at zero. Therefore, the objective of the present study was to evaluate commonly specified models using simulated datasets where the herd-level true prevalence was known. The specific purpose was to compare findings from models using hyperpriors to those using a simple beta distribution to model within-herd prevalence. A second objective was to investigate sources of error by varying characteristics of the simulated dataset. *Mycobacterium avium* subspecies *paratuberculosis* infection was used as an example for the baseline dataset. Data were simulated for 1000 herds across a range of herd-level true prevalence scenarios, and models were fitted using priors from recently published studies. The results demonstrated poor performance of these latent class models for diseases characterised by poor diagnostic test sensitivity and low within-herd true prevalence. All variations of the model appeared to be sensitive to the prior and tended to overestimate herd-level

46 true prevalence. Estimates were substantially improved in different infection scenarios
47 by increasing test sensitivity and within-herd true prevalence. The results of this study
48 raise questions about the accuracy of published estimates for the herd-level true
49 prevalence of paratuberculosis based on serological testing, using latent class analysis.
50 This study highlights the importance of conducting more rigorous sensitivity analyses
51 than have been carried out in previous analyses published to date.
52
53 **1. Introduction**

54 Prevalence is an important measurement of disease (or infection) occurrence.
55 Estimation of the true prevalence ($P_T$) within a population involves the application of a
56 diagnostic test to calculate apparent prevalence ($P_A$) and adjusting according to test
57 performance, sensitivity (Se) and specificity (Sp) (Rogan and Gladen, 1978).
58 However, there is often uncertainty regarding Se and Sp, and published values may
59 vary. Much of this variation can be attributed to differences among reference
60 populations and sampling strategies that have been used for the test validation
61 procedure (Greiner and Gardner, 2000). In addition, Se and Sp may vary according stage
62 of infection (Nielsen and Toft, 2008), prevalence (Brenner and Gefeller, 1997) and
63 between herds (Greiner and Gardner, 2000). It may therefore be unreasonable to
64 assume a fixed, constant, Se and Sp over different populations (Berkvens et al., 2006).
65 Consequently, the relationship between $P_T$ and $P_A$ can also be expected to vary between
66 populations.
67
68 The use of Bayesian latent class analysis for the estimation of herd (HTP) and animal-
69 level (ATP) true prevalence has become increasingly frequent in veterinary
70 epidemiology (Branscum et al., 2004). Using this approach, all parameters are

71 considered random variables that can be modelled using probability distributions.
72 Uncertainty and variability associated with estimates of test Se and Sp may therefore be
73 incorporated in the analysis. The resulting Bayesian posterior probability distribution
74 will provide inference on prevalence estimates, conditional on both currently observed
75 data and previous knowledge regarding the prevalence of infection.
76
77 To date, many of the studies that have estimated HTP using Bayesian latent class
78 analysis have examined cross sectional test data using models proposed by Hanson et
79 al. (2003). Using this approach, the number of animals testing positive in each herd is a
80 function of the within-herd ATP, and the performance of the test. However, to the
81 authors' knowledge this approach has not yet been evaluated using simulated data for
82 which the HTP is known and this is a fundamental step to assess model performance
83 when no gold standard is available.
84
85 Furthermore, using this approach, the ATP within infected herds is assumed to follow a
86 beta distribution, the parameters of which are estimated from hyperpriors. This method
87 aims to account for both the uncertainty and variability in within-herd ATP between
88 herds (Hanson et al., 2003). Hyperpriors are fitted as beta ($\mu$) and gamma ($\psi$)
89 distributions to model within-herd ATP in the form Beta ($\mu\psi$, $\psi(1-\mu)$) (Hanson et al.,
90 2003). However, McAloon et al. (2016) reported a potential issue when using
91 hyperpriors to estimate HTP of paratuberculosis in Irish dairy herds. This related to
92 issues fitting the hyperprior, and a tendency for the resulting beta distribution to invert
93 and become clustered at zero, which is counterintuitive given that it is used to model
94 true prevalence within infected herds, i.e. when prevalence is > 0 by definition. The
95 authors in that study therefore opted to use a simple beta distribution to model within-

herd true prevalence which incorporated both the uncertainty and the variability associated with the parameter, assuming an average within-herd ATP distribution over all herds. More recently, other authors have used a logit-normal distribution to model within-herd ATP of digital dermatitis infection in dairy cattle (Yang et al., 2017).

The consequences of using one approach to model within-herd ATP over another is not clear since HTP remains unknown. However, testing each method against simulated data with a known and fixed HTP would facilitate comparison of these methods whilst also providing an evaluation of the overall method. The first objective of this study therefore was to evaluate a Bayesian latent class analysis model for the estimation of HTP, using simulated datasets over a range of known HTPs and to compare findings from models using beta hyperpriors, logit-normal hyperpriors and those using a simple beta distribution to model within-herd ATP. Model inputs for the base model were based on estimation of paratuberculosis HTP as an example. Paratuberculosis infection is characterised by a poor test Se and generally low within-herd ATP. The second study objective was to investigate how different infection characteristics and test performance influence the accuracy of the model by increasing Se and within-herd ATP in the simulated datasets and in the priors for the corresponding estimating models.

## 2. **Materials and Methods**

### 2.1. Study population – data simulation

Table 1 shows the list of abbreviations used in the manuscript. Diagnostic test data were simulated for a range of known or actual HTP (aHTP), i.e. the proportion of herds with 1 or more infected cows. At each aHTP, data were simulated for 1000 herds as follows. The number of animals in each herd was drawn from a gamma distribution

(rounded to the nearest integer) which had been fitted to herd sizes from an earlier study (McAloon et al., 2016) using the "fitdistrplus" package in R (R Core Team, 2015), and each herd size was rounded to the nearest integer. The number of animals testing positive from each herd was then simulated with the following model;

$Npos_i \sim Binomial\ (P_{Ai}, herdsize_i)$

$P_{Ai} = Se \times ATP_i + (1-Sp) \times (1-ATP_i)$

$ATP_i = HTP_i \times CWHP_i$

$HTP_i \sim Bernoulli\ (aHTP)$

$CWHP_i \sim Beta(alpha_{CWHP}, beta_{CWHP})$

$Se \sim Beta(alpha_{Se}, beta_{Se})$

$Sp \sim Beta(alpha_{Sp}, beta_{Sp})$

$Herdsize_i \sim Gamma(S1, S2)$

Where $Npos_i$ was the number of test positive animals in the *i-th* herd; $Npos_i$ was drawn from a binomial distribution with a probability equal to the within-herd $P_{Ai}$, and n trials equal to the $herdsize_i$; $P_A$ was determined by the ATP in the i-th herd, and the test Se and Sp. $Herdsize_i$ was drawn from a gamma distribution rounded to the nearest integer. ATP was a combination of the HTP and the Conditional Within-Herd Prevalence (CWHP), defined as the within-herd ATP conditional on the herd being infected, i.e. when HTP > 0. HTP for the *i-th* herd was drawn from a Bernoulli distribution with a probability equal to the 'actual HTP' (aHTP). In the first instance, datasets were simulated across 3 different HTP scenarios: low HTP, with aHTPs of 0.10, 0.20, 0.30 and 0.40; medium HTP with aHTPs of 0.35, 0.45, 0.55 and 0.65; and high HTP, with aHTPs of 0.60, 0.70, 0.80 and 0.90. The use of these different HTP scenarios facilitated the use of low, medium and high priors to be used in the estimating model.

146 Datasets were simulated for a CWHP beta distribution with a mode of 0.05, and a 95$^{th}$
147 percentile of 0.15. Parameters of the input distributions are shown in Table 2 and R-
148 code for the simulation of the datasets is provided as Supplementary Material 1.

149 2.2. Prevalence estimation

150 The estimated Herd-level True Prevalence (eHTP) was then found using Bayesian latent
151 class analysis from these datasets. The model had the following model structure;

152 $Npos_i \sim$ binomial ($P_{Ai}$, $herdsize_i$)

153 $P_{Ai} = Se \times ATP_i + (1-Sp) \times (1-ATP_i)$

154 $ATP_i = HTP_i \times CWHP_i$

155 $HTP_i \sim$ Bernoulli (eHTP)

156 $Se \sim beta(alpha_{Se}, beta_{Se})$

157 $Sp \sim beta(alpha_{Sp}, beta_{Sp})$

158 CWHP was modelled in four different ways to compare the outcomes. The first model,
159 represented as BETA, used a simple beta prior distribution (McAloon et al., 2016)
160 whereas the second and third used beta hyperpriors from recently published studies,
161 called BETA-HYP1 (Verdugo et al., 2015) and BETA-HYP2 (Pozzato et al., 2011). These
162 distributions were in the form; Beta ($\mu\psi$, $\psi(1-\mu)$) where $\mu$ is a beta distribution used to
163 model the mean CWHP and $\psi$ is a gamma distribution used to model the variation
164 between herds. In this model structure, the degree of variation between herds is
165 inversely proportional to $\psi$ (Hanson et al., 2003); that is, with higher values of $\psi$, herds
166 will have more similar CWHP.
167
168 Although BETA-HYP1 and BETA-HYP2 were both originally used as priors to estimate
169 the prevalence of paratuberculosis, they were chosen to reflect the knowledge available
170 on those specific populations at a specific time. For this study, they were chosen as they

were relevant to paratuberculosis characteristics i.e. representing low CWHP, however, they also represented two variations of CWHP: one in which the prior for mean CWHP was quite precise, with moderate variation between herds (Verdugo et al., 2015) and the second in which the prior for mean CWHP was imprecise with a greater level of between-herd variation, i.e. with a higher mean ψ (15.8; Pozatto et al., 2011). The fourth model used a logit-normal distribution in the form $\text{logit}(CWHP_i) = \beta + \alpha_i$, where β is the logit-mean CWHP and $\alpha_i$ is a herd-level random effect modelled as a normal distribution with a mean of 0 and precision τ. This model structure was designated LOGIT-N. The form of each method is shown below and model priors are shown in Table 2.

*Model - BETA*

$CWHP_i \sim \text{beta}(\text{alpha}, \text{beta})$

*Model - BETA-HYP1/BETA-HYP2*

$CWHP_i \sim \text{beta}(\mu_i\psi_i, \psi_i(1-\mu_i))$

$\mu_i \sim \text{beta}(\text{alpha}, \text{beta})$

$\psi_i \sim \text{gamma}(S1,S2)$

*Model – LOGIT-N*

$\text{logit}(CWHP_i) = \beta + \alpha_i$

$\alpha_i \sim \text{norm}(0, 1/\tau)$

### 2.3. Sensitivity analysis

Sensitivity analysis was conducted by simulating and analysing a number of scenarios.

#### 2.3.1. eHTP prior

In each case, aHTPs were simulated across 3 different HTP scenarios (low, medium and high) as described above. For each of these scenarios, two different eHTP priors were trialled: firstly, a uniform beta(1,1) distribution was used as the prior for eHTP. Next, a beta prior which corresponded to the HTP scenario being simulated was also trialled. In the low HTP scenario, a beta prior with a mode of 0.25 was used, in the medium HTP scenario, a beta prior with a mode of 0.50 was used, and in the high HTP scenario a beta prior with a mode of 0.75 was used (Table 2).

#### 2.3.2. CWHP simulation method

In the base dataset, CWHP was simulated using a simple beta distribution. To assess the sensitivity of this method to the method used to simulate the data, alternative datasets were simulated in which CWHP was modelled using exactly the same model structure and inputs as the analytical model used for the estimation. For example, when assessing the accuracy of BETA-HYP1, this model was trialled on a dataset in which CWHP was simulated using a simple beta distribution, and a second dataset in which CWHP was modelled using the same model structure as the analytical model. In each case $\mu$ and $\psi$ were specified as distributions for the overall population. The CWHP for the *i-th* herd was then simulated by first drawing separately from these two distributions. These drawn values were used to generate parameters for a beta distribution, from which a single value was simulated as the CWHP of the herd. The datasets generated using the simple beta distribution and the dataset simulated according to the form of the estimating model were designated “Simple” and “Model Form” datasets respectively. The same approach was taken for BETA-HYP2 and LOGIT-N.

#### 2.3.3. Test and disease characteristics

For the second objective, we investigated how the accuracy of the prevalence estimates changed according to CWHP and test performance. The steps above were repeated under alternative infection scenarios with medium (mode, 0.5, 95% less than 0.6) and high (mode 0.8, 95% greater than 0.7) test Se; and for medium and high CWHP. For the CWHP sensitivity analysis, the distributions dictating the variability between herds, i.e. the gamma components for BETA-HYP1, BETA-HYP2 and LOGIT-N, were maintained from the base model, and only the parameters dictating the mean of the overall distribution were varied, i.e. the beta distributions for BETA-HYP1 and BETA-HYP2 and the normal distribution for LOGIT-N (Table 2).

Models were implemented in WinBUGS 4.3.1 (Lunn et al., 2000) with the first 5,000 iterations discarded as burn in and 15,000 iterations used for posterior inference. Convergence was assessed by visual inspection of the time series trace plots and by running multiple (n = 3) chains from different starting values. In all cases, chains reached stationary distributions within 5,000 iterations. A number of models were also run for 100,000 iterations check for identifiability issues.

## 3. Results

Figure 1 shows the distributions of CWHP simulated from each of the model structures. BETA-HYP2 in particular demonstrates significant clustering at zero as occurs when the alpha parameter of the beta distribution is <1.

Figure 2 plots the range of aHTP against the estimated HTP (eHTP) for low, medium and high HTP scenarios. Four main conclusions can be drawn from these figures: 1, in general, models were poor at estimating aHTP; 2. this estimation was not substantially

245 improved by varying the method used to model CWHP in the analytical model; 3, using
246 exactly the same model structure to simulate CWHP as that used for the analytical
247 model did not improve estimates, in fact, in many cases it appeared to make the
248 estimates worse; and 4, the estimates tended to be quite sensitive to the HTP prior used,
249 particularly with high HTPs. In the low HTP scenario, all the models tended to
250 overestimate HTP, with the exception of the BETA model which underestimated
251 prevalence for HTPs of 0.3 and 0.4, regardless of the prior used. Similarly, in the
252 medium HTP scenario, all models with the exception of the BETA model overestimated
253 HTP. In the high HTP scenario, estimates tended to cluster close to the HTP prior when
254 this was used, leading to overestimation of lower HTPs and under estimation of the 0.8
255 and 0.9 HTPs.
256
257 Figures 3 and 4 show the effect of varying the diagnostic test to medium and high Se
258 respectively. In general, accuracy of estimates are improved considerably with
259 increasing Se across all of the methods used to model CWHP. Both figures show
260 substantially improved HTP estimates and a much-reduced sensitivity to the prior for
261 HTP. Overall, there is still a tendency for models to overestimate HTP, particularly
262 models BETA-HYP1 and BETA-HYP2 and this tendency is reduced as test Se is
263 increased. The accuracy of the models are substantially improved at higher aHTPs,
264 particularly in the simple dataset. In contrast to the base model, there appears to be a
265 small improvement in using the same model structure for the simulation.
266 Figures 5 and 6 show the effect of increasing CWHP on the accuracy of the model. In
267 general, estimates were improved relative to the base scenario. However, in the
268 medium CWHP scenario, some large positive deviations in eHTP relative to aHTP may
269 be observed. This appears to be particularly evident at low aHTPs in the BETA-HYP2

model and in the model form scenarios, which could be related to the fact that the CWHP distributions used to model this scenario include a large amount of between-herd variability in CWHP.

## 4. Discussion

The use of simulated data to assess and compare the effectiveness of mathematical models is a useful method of model evaluation that is commonly used within the field of genetics (Stephens and Donnelly, 2003; Wilson and Rannala, 2003; Faubet et al., 2007)) and has gained increasing popularity with the field of veterinary epidemiology (Denwood et al., 2010; Singleton and Breheny, 2016). Similarly, in veterinary epidemiology, the use of Bayesian models to estimate prevalence has also increased in recent years and is often used to estimate the prevalence of paratuberculosis, because of uncertainty around the performance of diagnostic tests (Liapi et al., 2011; Pozzato et al., 2011; Verdugo et al., 2015; McAloon et al., 2016). However, to the authors' knowledge this is the first study that has used simulated data to evaluate the overall accuracy of Bayesian latent class analysis for the estimation of HTP, and to evaluate the effect of varying components within the model, for example the use of hyperpriors for modelling CWHP.

This study raises substantive concerns about the effectiveness of conventional Bayesian latent models to estimate paratuberculosis HTP and this may apply to other infections or diseases with similar diagnostic test characteristics and where within-herd prevalence is often very low. Irrespective of the method used to model CWHP, our models tend to overestimate HTP. The HYP1, HYP2 and LOGIT-N models produced estimates with larger probability intervals, whereas the BETA model produced median

values that were closer to aHTP, but with much narrower probability intervals. There was little difference between the two hyperprior methods of modelling CWHP, however, HYP2 tended to produce less predictable estimates in response to increasing aHTP in comparison to HYP1 (Figure 5).

Importantly, when used in the paratuberculosis scenario, all models appeared to be overly sensitive to the prior used for HTP, particularly when a high HTP prior was used. Interestingly, in the worked example in Branscum (2004), we note that the median and 95$^{th}$ percentile of the posterior estimate for HTP (0.58, 0.83 respectively) were also notably close to the median and 95$^{th}$ percentile from the prior distribution (0.59, 0.85 respectively). Similarly, in published examples of the method, Pozatto (2012) found that the HTP (median, 95% credible intervals) in 2 regions in Italy was 0.70, 0.50-0.87 and 0.71, 0.54 – 0.87, whilst the prior distribution used for HTP in this study was 0.69, 0.50-0.84. Liapi et al., (2011) used a prior of 0.65 with a 5$^{th}$ percentile of 0.40 and found a posterior estimate of 0.61 and 0.42 respectively. In Bayesian analyses, when posterior estimates closely reflect prior distributions, there is cause for concern that the data are having little impact on the results, which suggests models may not be appropriately specified. A greater difference between prior and posterior estimates was found in Verdugo et al. (2015) who reported posterior estimates for HTP of 0.92 (0.87-0.96), 0.78 (0.74-0.83) and 0.75 (0.71-0.78) with a prior of 0.86 (0.59 – 0.95), however this model used a different approach which allowed for an age-specific sensitivity for each animal which were higher than the Se estimates used in other analyses. This study was based on a larger sample size, however, our analyses have shown that the problems identified with this method cannot be overcome by increasing sample size (data not shown).

Figures 5 and 6 show large deviations of eHTP relative to aHTP at specific aHTP values, for example in the BETA-HYP2 model on the Model Form dataset, under the low HTP scenario (Figure 5). In these cases, the posterior distribution for eHTP was very high relative to the aHTP, whereas the posterior estimate for Se was very low, approaching zero. Repeat analysis with multiple chains showed stability of separate chains at two different parameter spaces suggesting a problem with model identifiability. These issues were not resolved by running the model for more (n=100,000) iterations or by reducing the uncertainty around the Se prior but could be 'fixed' by varying the initial starting values. In practice it may not be possible to know what the 'true' model is, therefore for future studies, it is particularly important that multiple chains are run from a variety of initial values, to check for identifiability issues. In addition, examination and reporting of the posterior distributions for the rest of the parameters in the model is also recommended, including those parameters that are not specifically of interest.

Studies using simulation to assess model accuracy often generate a reasonably large number of datasets from a particular model with particular parameters. Each of these datasets is analysed, and the results used to examine the performance of the estimation method. For example, Singleton et al. (2016) used simulated data to assess the utility of a non-linear hierarchical model applied to experimental infection data. Three sample sizes were chosen, and 5,000 datasets generated for each set of parameters with each dataset analysed by the proposed model. In the case of our study, the outcome of interest at each aHTP was a known point prevalence which would not change if additional datasets were generated. For each aHTP however, 1,000 herds were

345 simulated for each set of parameter values, representing the replicated datasets to
346 assess the method.
347
348 The use of hyperpriors to model within-herd ATP is commonly advocated in the use of
349 latent class estimation of HTP. Using this method, hyperpriors are fitted as beta (μ) and
350 gamma (ψ) distributions to model within-herd ATP in the form Beta (μψ, ψ(1-μ))
351 (Hanson et al., 2003). The potential advantage of this method is that it facilitates the
352 incorporation of both uncertainty regarding the parameter as well as the between-herd
353 variability. The distributions are fitted through the elicitation of expert opinion, who are
354 asked to specify the mean and confidence intervals of the within-herd ATP across herds,
355 which is fitted as a beta distribution (μ). Then, conditional on the mean, experts are
356 asked to specify the value below which they are 95% sure that 90% of the within-herd
357 ATP are below. These values are then used to fit the gamma distribution (ψ). However,
358 whilst this method has obvious theoretical advantages, we argue that the data required
359 from expert elicitation may be restrictively complex. Furthermore, McAloon et al.
360 (2016) highlighted inconsistencies in published literature between values elicited from
361 experts and those same percentiles based on simulation of the hyperprior distributions.
362 Finally, given that within this method, distributions are fitted conditional on a mean,
363 rather than mode, the distribution often becomes inverted, and very often the median
364 prevalence within infected herds may be less than 0.01. This is potentially problematic
365 with small to medium herd sizes as herds may be deemed infected yet have less than 1
366 infected cow in the herd. We hypothesised that this may result in overestimation of
367 HTP. The present study seems to suggest that the use of beta hyperpriors does appear
368 to overestimate the HTP more so than the BETA or LOGIT-N models. This

overestimation is particularly evident with priors that incorporate increased variability in CWP, for example the BETA-HYP1 model.

A possible explanation for this finding is that this method fits the 90$^{th}$ percentile conditional on a fixed mean. However, given a beta distribution with a fixed mean, increasing the variance in order to increase the 90$^{th}$ percentile leads to a shift in the median in the opposite direction creating an increasing skewed distribution. If the mean is low as is the case in paratuberculosis, the median moves very close to 0 as the alpha parameter becomes < 1. With very low CWHP, the probability of herds being infected with an AP of 0 increases potentially leading to this herd being "infected" across more iterations. In contrast, the LOGIT-N method, facilitates increased variation but still retains a distribution shape that is possibly more reflective of the likely distribution (Figure 1). However, the overall effect of this problem with the method of modelling CWHP was relatively minor when compared with the problems associated with the overall use of the model to estimate HTP of paratuberculosis, with relatively poor Se and low CWHP. Increasing the Se to 0.5 and 0.8 led to increases in the accuracy of the estimates. Similarly, increasing the mode of the distribution used to model mean CWHP to 0.3 and 0.7 also led to increased accuracy of the estimates and a decreased sensitivity to the HTP prior used, across all of the models used. Therefore, these models may be reasonably accurate when used to estimate prevalence for infections or diseases with poor Se *or* low CWHP but not when both of these are present.

In addition, it is important to note that during the simulation stage of this study, the "design" aHTP used to generate the simulated dataset may have differed from the actual proportion of herds in the simulated dataset with one or more infected animals. This

occurred because herds were first simulated as infected by drawing from a Bernoulli distribution with a probability equal to the aHTP. Within those herds deemed infected, the number of infected individuals was then drawn from a binomial distribution with a probability equal to CWHP drawn for that herd. However, with moderate herd sizes and low CWHP, the probability of drawing zero infected individuals in an "infected" herd is >0 and increases with decreasing CWHP. Within the low CWHP datasets, the difference between the design and actual datasets was greatest for the BETA-HYP1 and BETA-HYP2 models compared to the BETA and LOGIT-N models, probably because of the greater tendency for this model structure to become clustered at zero. All of the models in general tended to overestimate aHTP, and aHTP may be an overestimate of the actual proportion of infected herds.

## 5. Conclusion

Our results suggest poor accuracy of commonly specified Bayesian latent class models for paratuberculosis herd-level true prevalence estimation. All variations of the model appeared to be sensitive to the prior and tended to overestimate herd-level true prevalence, raising questions about whether previous estimates of paratuberculosis HTP reported in the literature may be inaccurate. Estimates were substantially improved in different infection scenarios by increasing test sensitivity and within-herd true prevalence. This study highlights the importance of conducting more rigorous sensitivity analyses than have been carried out in previous analyses published to date. In addition, we advocate increased use of simulation as an initial stage in conducting future analyses and also suggest that new model methodologies be explored, to determine whether alternative approaches might perform better than conventional latent class models.

**Acknowledgements**

Conor McAloon is supported by a UCD Wellcome Institutional Strategic Support Fund, which was financed jointly by University College Dublin and the SFI-HRB-Wellcome Biomedical Research Partnership (ref 204844/Z/16/Z).

## 6. References


Berkvens, D., Speybroeck, N., Praet, N., Adel, A., Lesaffre, E., 2006. Estimating disease prevalence in a Bayesian framework using probabilistic constraints. Epidemiology 17, 145-153.

Branscum, A., Gardner, I., Johnson, W., 2004. Bayesian modeling of animal-and herd-level prevalences. Preventive veterinary medicine 66, 101-112.

Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Statistics in medicine 16, 981-991.

Denwood, M.J., Reid, S.W.J., Love, S., Nielsen, M.K., Matthews, L., McKendrick, I.J., Innocent, G.T., 2010. Comparison of three alternative methods for analysis of equine faecal egg count reduction test data. Preventive veterinary medicine, 93, 316-323.

Dorshorst, N.C., Collins, M.T., Lombard, J.E., 2006. Decision analysis model for paratuberculosis control in commercial dairy herds. Preventive veterinary medicine 75, 92-122.

Espejo, L., Godden, S., Hartmann, W., Wells, S., 2012. Reduction in incidence of Johne's disease associated with implementation of a disease control program in Minnesota demonstration herds. Journal of dairy science 95, 4141-4152.

Faubet, P., Waples, R.S., Gaggiotti, O.E., 2007. Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. Molecular ecology, 16, 1149-1166.

Ferrouillet, C., Wells, S., Hartmann, W., Godden, S., Carrier, J., 2009. Decrease of Johne's disease prevalence and incidence in six Minnesota, USA, dairy cattle herds on a long-term management program. Preventive veterinary medicine 88, 128-137.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Preventive veterinary medicine 45, 3-22.

449 Hanson, T., Johnson, W.O., Gardner, I.A., 2003. Hierarchical models for estimating herd
450 prevalence and test accuracy in the absence of a gold standard. Journal of Agricultural,
451 Biological, and Environmental Statistics 8, 223-239.
452 Johnson-Ifearulundu, Y., Kaneene, J., Sprecher, D., Gardiner, J., Lloyd, J., 2000. The effect
453 of subclinical Mycobacterium paratuberculosis infection on days open in Michigan, USA,
454 dairy cows. Preventive veterinary medicine 46, 171-181.
455 Liapi, M., Leontides, L., Kostoulas, P., Botsaris, G., Iacovou, Y., Rees, C., Georgiou, K.,
456 Smith, G.C., Naseby, D.C., 2011. Bayesian estimation of the true prevalence of
457 Mycobacterium avium subsp. paratuberculosis infection in Cypriot dairy sheep and goat
458 flocks. Small Ruminant Research 95, 174-178.
459 Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS — a Bayesian
460 modelling framework: concepts, structure, and extensibility. Statistics and Computing,
461 10:325–337.
462 McAloon, C.G., Doherty, M.L., Whyte, P., O'Grady, L., More, S.J., Messam, L.L.M., Good, M.,
463 Mullowney, P., Strain, S., Green, M.J., 2016. Bayesian estimation of prevalence of
464 paratuberculosis in dairy herds enrolled in a voluntary Johne's Disease Control
465 Programme in Ireland. Preventive veterinary medicine 128, 95-100.
466 Nielsen, S.S., Toft, N., 2008. Ante mortem diagnosis of paratuberculosis: a review of
467 accuracies of ELISA, interferon-γ assay and faecal culture techniques. Veterinary
468 microbiology 129, 217-235.
469 Pillars, R., Grooms, D., Gardiner, J., Kaneene, J., 2011. Association between risk-
470 assessment scores and individual-cow Johne's disease-test status over time on seven
471 Michigan, USA dairy herds. Preventive veterinary medicine 98, 10-18.

Pozzato, N., Capello, K., Comin, A., Toft, N., Nielsen, S.S., Vicenzoni, G., Arrigoni, N., 2011. Prevalence of paratuberculosis infection in dairy cattle in Northern Italy. Preventive veterinary medicine 102, 83-86.

Rogan, W.J., Gladen, B., 1978. Estimating prevalence from the results of a screening test. American journal of epidemiology 107, 71-76.

Singleton, M.D., Breheny, P.J., 2016. Nonlinear hierarchical modeling of experimental infection data. Preventive veterinary medicine, 130, 129-136.

Stephens, M., Donnelly, P., 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. The American journal of human genetics, 73, 1162-1169.

Verdugo, C., Toft, N., Nielsen, S.S., 2015. Within-and between-herd prevalence variation of Mycobacterium avium subsp. paratuberculosis infection among control programme herds in Denmark (2011–2013). Preventive veterinary medicine 121, 282-287.

Wilson, G.A., Rannala, B., 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics, 163, 1177-1191.

Yang, D.A., Heuer, C., Laven, R., Vink, W.D., Chesterton, R.N., 2017. Farm and cow-level prevalence of bovine digital dermatitis on dairy farms in Taranaki, New Zealand. New Zealand veterinary journal, 65(5), pp.252-256.