

A modelling framework for the prediction of the herd-level probability of infection from longitudinal data

Aurélien Madouasse, Mathilde Mercat, Annika van Roon, David Graham, Maria Guelbenzu, Inge Santman Berends, Gerdien van Schaik, Mirjam Nielen, Jenny Frössling, Estelle Ågren, Roger Humphry, Jude Eze, George Gunn, Madeleine Henry, Jörn Gethmann, Simon J. More, Christine Fourichon

<https://doi.org/10.1101/2020.07.10.197426> version
<https://www.biorxiv.org/content/10.1101/2020.07.10.197426v3>
Submitted by Aurélien Madouasse 2020-07-23 08:13

Abstract

For many infectious diseases of farm animals, there exist collective control programmes (CPs) that rely on the application of diagnostic testing at regular time intervals for the identification of infected animals or herds. The diversity of these CPs complicates the trade of animals between regions or countries because the definition of freedom from infection differs from one CP to another. In this paper, we describe a statistical model for the prediction of herd level probabilities of infection from longitudinal data collected as part of CPs against infectious diseases of cattle. The model was applied to data collected as part of a CP against infections by the bovine viral diarrhoea virus (BVDV) in Loire-Atlantique, France. The model represents infection as a herd latent status with a monthly dynamics. This latent status determines test results through test sensitivity and test specificity. The probability of becoming status positive between consecutive months is modelled as a function of risk factors (when available) using logistic regression. Modelling is performed in a Bayesian framework. Prior distributions need to be provided for the sensitivities and specificities of the different tests used, for the probability of remaining status positive between months as well as for the probability of becoming positive between months. When risk factors are available, prior distributions need to be provided for the coefficients of the logistic regression in place of the prior for the probability of becoming positive. From these prior distributions and from the longitudinal data, the model returns posterior probability distributions for being status positive in all herds on the current months. Data from the previous months are used for parameter estimation. The impact of using different prior distributions and model settings on parameter estimation was evaluated using the data. The main advantage of this model is its ability to predict a probability of being status positive on a month from inputs that can vary in terms of nature of test, frequency of testing and risk factor availability. The main challenge in applying the model to the BVDV CP data was in identifying prior distributions, especially for test characteristics, that corresponded to the latent status of interest, i.e. herds with at least one persistently infected (PI) animal. The model is available on Github as an R package (<https://github.com/AurMad/STOCfree>).

Keywords: Bayesian latent class model; disease surveillance; cattle

Round #1

Author's Reply:

by Rowland Raymond Kao, 2020-12-01 21:03

Recommendation for revision of "A modelling framework for the prediction of the herd-level probability of infection from longitudinal data"

Two reviews have been received for this preprint. Both reviewers have highlighted the importance of the problem - i.e. the estimation of herd "freedom from infection" status based on a series of partial observations with imperfect testing, which is particularly a problem where disease incidence is long, and the disease itself with long and variable latent periods. This is the case for BVD (the disease observed here) but also problematic long duration bacterial diseases such as paratuberculosis and bovine Tuberculosis. The development of a general framework, described here, is therefore welcome. Both reviewers have indicated the importance of the problem and relevance of the approach. Both have also been very positive about the execution of the approach, but both have also highlighted methodological issues that, while they do not fundamentally compromise this analysis, do require addressing.

A particular issue is the choice and use of priors, highlighted by both reviewers though with different perspectives. It also seems to me that the choice of prior is one area where other adopters of the framework would have to exercise the greatest individual judgement (also, as indicated in their discussion of this from lines 635 onwards, they are very aware of this) and therefore the reviewers comments require considerable attention. I also fully agree with one reviewer on the justification of the use of narrow priors - it suggests possible issues with the model specification itself or with the data, and is not necessarily an issue of the prior.

While some of the other reviewer comments are outside my expertise, they all seem sensible and should be responded to by the authors.

We would like to thank all 3 reviewers for their useful and constructive comments, which have helped us to improve the content of the paper.

Apologies for the time it has taken us to reply. Following the reviewers' comments and various feedback we have received, we have made some major changes to the model implementation which have required a lot of recoding.

First, following one of the reviewer's comment, it is now possible to specify priors for the status dynamics as normal distributions on the logit scale. This makes more sense. It is now easier to compare models with and without risk factors.

Second, initially, in order to speed up computation in JAGS, we have switched to the runjags package (instead of rjags) which allows using multiple cores and makes the model simpler to use, explore and modify. This has significantly reduced the time required to run our models.

Third, because of the recurrent convergence problems when priors are too wide in the JAGS model, we have implemented the model in Stan. This required some work on our part because Stan does not work with latent discrete parameters. We have adapted an implementation of the forward algorithm developed for another application (<https://github.com/luisdamiano/stancon18>). The Stan implementation works much better: the convergence problems are not as important and the model runs much faster, notably because we need less iterations because we have less autocorrelation in our posterior samples. So, we have decided to compare the 2 implementations in the paper.

Lastly, we have included summary and plotting functions in the package which should help exploring the model results in a simpler way.

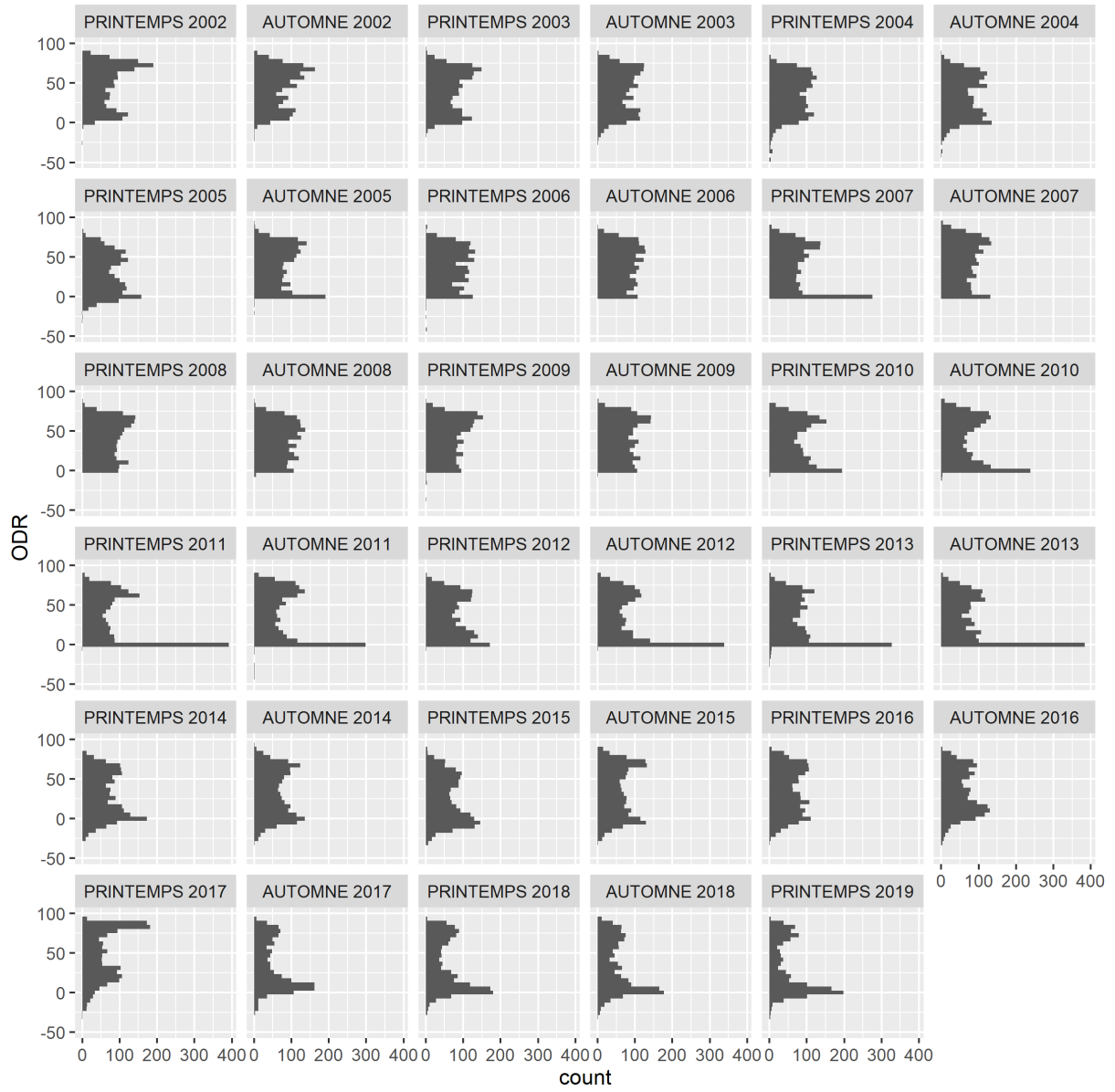
Because the model does not require such narrow priors as in the previous version, we have removed the part on confirmatory testing. This part was needed for the JAGS model to converge, but because the confirmatory test results were randomly sampled, it was not entirely satisfactory. We now have chosen to compare the Stan and JAGS implementation of our model.

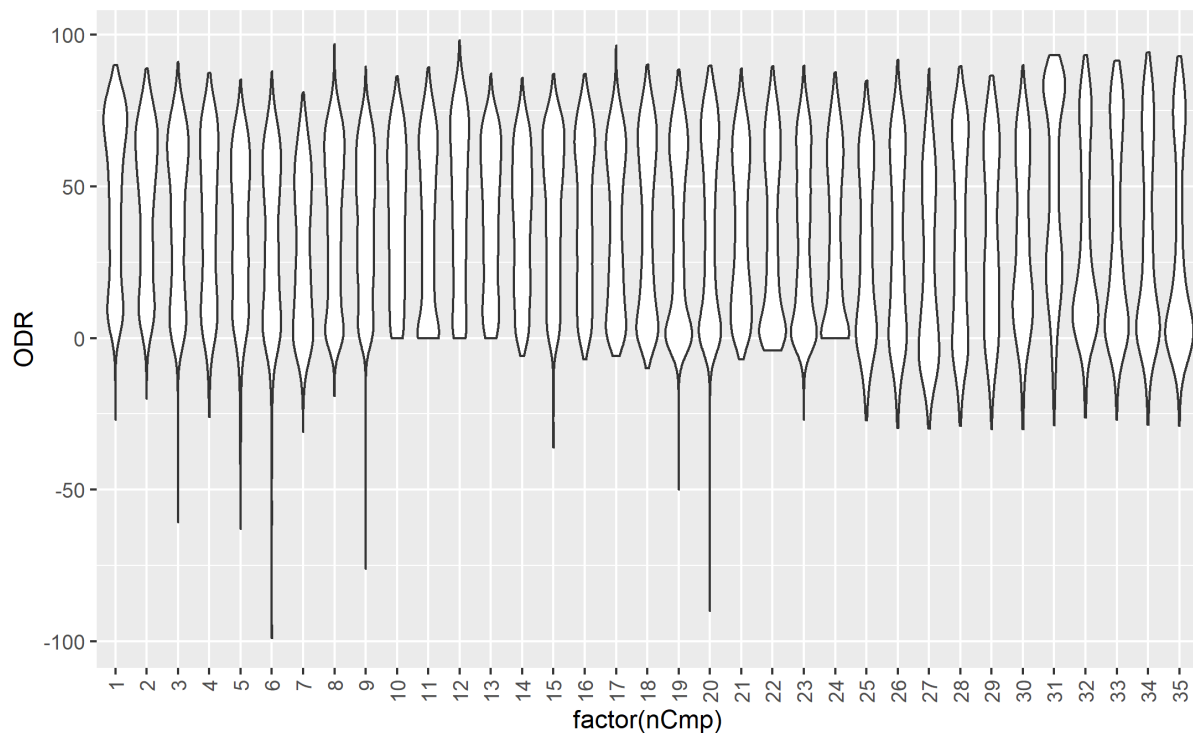
Problem of determining the parameters of the prior distributions for test characteristics

We address this problem here as it was raised by all reviewers. This is less critical than in the first version because we used much wider priors this time. This was possible because the Stan model can cope much better than JAGS with it. As a consequence, we have moved the mixture model part to a supplementary material file.

We agree that this is not good practice to use the same data for determining the priors and for analysis. The problem we faced was that we had no published data or expert opinion to rely on, since these test data are used in an empirical manner in the programme they originate from. So we decided to model the observed data as a mixture of normal distributions and to derive sensitivity and specificity from these fitted distributions. An alternative could have been to keep the ODR data as continuous and to model them as mixture of normal distributions directly in the Bayesian model. This could be useful in the future, but this is not what the users of this model would expect, and was therefore not considered.

To overcome this problem, we have used test data collected outside of the study period. We had data on all tests performed between spring 2002 and spring 2019. The difficulty this time was that before the study period, the lab doing the analyses often truncated the data at 0. Mixture models would not work with such data. Then in spring 2017, the lab used a more sensitive, which was later abandoned. This is visible in the 2 Figures below which represent the distributions of the optical density ratios in the different testing campaigns.





For our mixture models, we therefore chose to use the data from the 4 campaigns between autumn 2017 and spring 2019. We assumed that the data collected during the study periods and these 4 testing campaigns followed the same distributions, conditionally on seropositivity, since the same test was used. The distribution was still bi-modal, with a higher proportion of the observations on the left hand-side part. The somewhat imperfect fit of the mixture to the data was even more visible with this dataset than with the one used in the previous version of the paper. Therefore, we chose to use a mixture of 4 normal distributions and consider the 2 distributions on the left to represent sero-negatives and the 2 distributions on the right to represent the sero-positives. As in the previous version, we used a cut-off of 35 which is associated with a sensitivity of 0.978 and a specificity of 0.95.

I also include here some comments from a 3rd reviewer who did not feel it sufficiently within their expertise to offer a full review:

"The submitted manuscript aims to predict herd status using all prior information over time. This is an interesting new approach and worth being published. However, author imply in the introduction that they consider the number of animals tested (a frequent feature of herd testing) whereas they don't consider this option. Only tank milk testing is included.

In the introduction, we present the number of animals sampled as something that can vary between control programmes, but this was not investigated in our study. Our package offers the possibility to include test data as a number of positives out of a number of tested (only in the JAGS version). But we feel that this is not ready for publications because, as it is, it makes the hypothesis that animals are randomly sampled from a herd and that all the sampled animals have the same probability of infection. This is usually not the case, especially with BVD where young calves are more likely to be infected and various age groups can be tested in different programmes. Our current recommendation would be to consider the outcome of

testing several animals in a month as a single herd level test result with a single pair of sensitivity / specificity. This is an area that would warrant further research.

Moreover, I believe there is a mistake in formula 15 for the probability of infection for test-negative herds (I think this should read: $(1-T+).[Sp.(1-pS+) / (Sp.(1-pS+) + (1-Se).pS+)$ i.e. the negative predictive value for negative test results). Having said that, my advice is to consult a statistician for quality checking of the code. I recommend consulting my colleagues Prof Geoff Jones (g.jones@massey.ac.nz) or Prof Wesley Johnson (wjohnson@ics.uci.edu).

Thank you for pointing this out. We corrected the mistake in the paper. The corresponding code was correct. Regarding quality checking, we have involved Nils Toft in our work. Nils has done extensive work on Bayesian latent class modelling. He has checked the model and has found no major issue with it.

The text requires an overhaul for clarity and typos.

The text has been reviewed by the native English co-authors. We hope that the current content is clear.

The manuscript is definitely worth reviewing as it is topical at a time of emerging infectious diseases, so it would likely attract citations. I cannot review this submission myself as it would need considerable more time and more statistical expertise than I can offer."

Overall, this is likely to be a very good contribution to the literature and I look forward to seeing a revision.

Reviews

Reviewed by Arata Hidano, 2020-09-21 03:14

Review Report

21st September 2020

Madouasse et al 2020. A modelling framework for the prediction of the herd-level probability of infection from longitudinal data

Overall comments:

Difference in surveillance and disease control intensities makes the interpretation of herd status difficult, particularly for diseases such as BVD and Johne's that have potentially long latent periods. The authors used a Bayesian framework to integrate various sources of information to calculate the probability of herds having at least one BVD persistently infected (PI) animal. I appreciate the motivation behind this study because the ability to predict herd disease status in near real-time may provide useful information that guide relevant stakeholders to make decisions on trading and disease controls. Providing the code and R package is excellent and this will certainly facilitate the discussions on this topic within the community. Therefore, overall, I would very like to see this work to be shared widely and this kind of framework to be applied to many data held in industries and governments.

I have several questions that I would like the authors to address and/or clarify as follows.

1. Prior for test accuracy

I have a few comments on the prior for the test accuracy. Which years of testing results (9725 observations) were used to estimate the prior for the bulk milk ELISA sensitivity and specificity? I understand that Bayesian modelling was done for data between 2014 and 2016. Because prior should not be elicited from the same data that is analysed (by definition of prior), please clarify whether test accuracy prior was obtained from data before 2014.

We agree that it is not a good practice. We have used a different period to estimate the priors for sensitivity and specificity. See the full justification in our general comments. With the Stan implementation of the model, the prior distributions on test characteristics do not have to be as narrow as with the Stan

As I initially read this manuscript, I wondered why the prior for the test sensitivity was extremely narrow (5th percentile 94.6 and 95th percentile 95.5). This question was answered by Line 640 onwards suggesting that there was non-identifiability issue when wider prior was used. Using a narrow prior itself is not an issue as long as this can be justified; however, non-identifiability indicates the lack of information contained in the data and this would not really justify the use of narrower prior. Having said that, I understand the parameter estimation was probably not the interest of this study. I would therefore suggest removing Line 640 onwards because I'm afraid that these sentences may give an impression to users who are less familiar with Bayesian analysis that prior can be adjusted flexibly so that convergence can be achieved. Instead, please state clearly that prior for test accuracies should be drawn independently from the data using e.g. literature, expert opinions, previous data, or diffuse priors.

This was indeed a major problem with the previous version of the model. The Stan version has made things much easier. We have used much wider priors in this version. The estimation of sensitivity and specificity from the mixture model serves as a rough guide to get plausible values.

2. Logistic regression

It is excellent that the logistic model dynamically predicted probabilities of the latent status changing from negative to positive, accounting for risk factors as well as uncertainties in the test outcome. It is a trivial comment; however, it seems that the latent status was defined as presence/absence of at least one PI animal (Line 206) but explanations about the logistic model often said it was predicting the probability of infection (e.g. Line 339). Please use consistent definitions for the latent status throughout the manuscript.

This has been changed to *The model predicts herd-level probabilities of being latent status positive on the last month.* (L321 of the current version)

3. Confirmatory test

The manuscript mentioned that the results of confirmatory test were randomly generated once herds tested positive by the routine test. Was the data of confirmatory test (e.g. when, which herds were tested and their results) available? If so I wonder why this data was not explicitly integrated into the modelling framework i.e. use the observed positive/negative test results to estimate parameters rather than generating random results. Adding the observed second test results may have solved the non-identifiability issue. I may have missed something, so please clarify. Also please add information about the prior distribution used for the test accuracy of the confirmatory test (Line 466). Or did it assume 100% DSe and DSp (Line 567)?

This was indeed added to resolve the non-identifiability issue. We did not have the data to do that, so we generated confirmatory tests after a positive test. The results of these confirmatory tests were generated at random. Because the Stan version of the model resolves the identifiability problem, we have removed this part of the paper.

4. Model comparison

It was not immediately clear to me what the motivation was behind for developing 4 different models. Was the intention to show that this framework can handle different information source and demonstrate how it can be done? Perhaps if a sentence could be added somewhere that clearly states the objective, it would make it easier to read the text.

We have re-organised the analyses. We hope this version is clearer.

Minor comments:

Line 139: For the estimation to work, the same tests should be used in different populations... I assume this sentence talks about frequentist framework; these assumptions can be relaxed in various ways in Bayesian framework. Please clarify this in the text.

We have added *although some of these assumptions can be relaxed in a Bayesian framework.* (L 147)

Line 165: ...although not with longitudinal data from multiple epidemiological units such as herds. I'm aware that Simon Spencer from Warwick has been working on this field and there are some applications of HMM for veterinary data (although precisely speaking they don't have many units). I would suggest adding some previous works e.g. (Touloupou et al., 2020).
Touloupou, P., Finkenstädt, B., Spencer, S.E.F., 2020. Scalable Bayesian Inference for Coupled Hidden Markov and Semi-Markov Models. *Journal of Computational and Graphical Statistics* 29, 238–249.
<https://doi.org/10.1080/10618600.2019.1654880>

This reference has been added.

Line 246: The status on the first time step (S_1^+)...
It was slightly confusing that S^+ was used to represent the latent status itself (regardless being positive or negative), while it also sometimes represent 'positive' status only e.g. formula in Line 248, which is showing the probability of being status positive. Perhaps re-writing as $p(S_1^+=+ve)$ or similar throughout the manuscript would avoid this confusion.

This was indeed confusing. We have replaced the S^+ by $S = 1$ and $p(S = 1)$ by π . We hope the different equations are easier to read.

Line 308 'specificity of interest'
Did it mean sensitivity?

Yes. We have removed this paragraph as suggested by the other reviewer.

Line 712: dynamics contributed to the definition...
I would suggest using 'estimation' rather than 'definition'.

This has been changed. (L714)

Line 762: 95% credibility [interval] for the association between local seroprevalence and new infection included 0 and this variable was therefore not included.

This is just a trivial comment and out of curiosity; I wonder the inclusion of this variable may have improved the estimation of other parameters? Did you record Bayes Factor with/without this variable, or credibility interval of this variable?

We did not investigate this further. From what I can remember, it did not have a major impact on the other model estimates.

Reviewed by anonymous reviewer, 2020-11-30 19:00

The authors present an MCMC modelling framework intended for use with serial observations of disease status, with a particular focus on the use of imperfect diagnostic tests. I find the motivation to be clear and compelling, the methodological approach to be sound, and the manuscript generally well written. However, I do have a number of comments that I think may improve the manuscript:

1. I am concerned about the apparent re-use of the same data for both informing the priors for Se and Sp (lines 454-456; 488-490) and presumably also contributing to the likelihood function of the fitted model, presuming that these same data points are used for model fitting? In that case the data will have a double contribution to the posterior, which is not a good idea. Can you either reassure the reader that the same data points are not used twice (either directly or indirectly) or base the priors on other information that is not already used in the model?

We have used a different dataset to estimate sensitivity and specificity this time. See our general comment above. Because the Stan implementation of the model does not require priors that are as narrow as the JAGS implementation, we find this point to be less critical and have moved the details of this analysis to supplementary material (see S1).

2. There seem to be considerable challenges with convergence with the model, presumably resulting from high autocorrelation. However, the brief mention of MCMC convergence diagnostics on lines 468-473 does not specifically mention the degree of autocorrelation - was it not a problem after all? Also, can you give details on the effective sample size based on your 4 chains of 5000 correlated samples with thin of 20? Even if correlation in the 1000 samples is extremely low I would still worry about a minimum effective sample size of 400 being reached for all parameters.

The package has been updated with a summary function that returns the effective sample size for each model parameter. We have added these summaries as supplementary materials (see S2). The effective sample sizes were usually high, especially with the Stan version of the models.

3. A mixture of normals is assumed for the observed OD values, but (to me) figure 5 indicates a somewhat imperfect fit to the data: the left peak seems right-skewed and the right peak seems left-skewed, i.e. there are more observations in the middle than

expected. This may have the effect of exaggerating the diagnostic capability of the test, as the "crossover" between the assumed normal distributions would be a lot less than that in the real data. How might this have affected your results?

We now have a better fit with 4 instead of 2 normals. Because we can now use wider prior distributions, we just need rough guides.

4. Model 1 uses a beta distribution for T1, but Models 2-4 use a logistic regression for T1 (Table 3). However, you have not fully considered the effect of the priors being on different scales (i.e. probability scale vs logit scale) - an excellent example of the problems this can give can be found in "Gelman A, Simpson D, Betancourt M. The prior can often only be understood in the context of the likelihood. Entropy. 2017;19:1-13.". Please re-run Model 1 using an intercept-only model with prior equivalent to the other models so that we can be sure the difference between the models is really due to inclusion of the risk factor, and not due to an indirect effect of the prior for the "average" probability being on different scales. Also, I would strongly advise to center the predictors to zero so that the intercepts can be directly compared between models - this should also have a beneficial effect on convergence.

The model has been recoded to allow modelling the probability of new infection (τ_1) on the logit scale, regardless of the presence of risk factors. The paper has been updated accordingly.

We agree that centering predictors can help interpretation. We only included one predictor in the model, the log of the number of cattle introduced + 1. The way this predictor was coded, the intercept was the logit probability of new infection when 0 cattle were introduced. We think that it is easier to determine a prior distribution for the intercept with our covariate coded this way than when considering the introduction of an "average" number of cattle.

5. I found figures 7, 9 and 10 to be difficult to interpret due to overlapping densities - is there some way of presenting these results in a different manner, perhaps using a caterpillar plot / horizontal boxplot, or maybe an ECDF plot?

The plots have been updated accordingly. We hope they are easier to read in this version.

I also have a number of more minor comments:

- Line 175: I agree it that missing observations are workable within the framework, but should you mention that the fewer missing observations the better (both in terms of convergence and precision of inference)?

The sentence has been updated: *Test results do not have to be available at every time step for the model to work, although the estimation will be more accurate with a large number of test results.* (See L181 of the current version)

- Line 195: I think the preferred reference for JAGS is "1. Plummer M. JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling JAGS : Just Another Gibbs Sampler. Proc 3rd Int Work Distrib Stat Comput (DSC 2003)

[Internet]. 2003. p. March 20–22, Vienna, Austria. ISSN 1609-395X. Available from: <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf>"

The reference has been updated

- Line 217: The observed distributions are never identical; maybe "are drawn from the same distribution" or "have close to the same distribution".

Changed to: *For a variable, if all the MCMC draws from the different chains are drawn from the same distribution ...* (See L229 of the current version)

- Equation 4: Obviously the effect of equation 4 is that $P(s_i+)$ is EITHER equal to T1 or T2 conditional on the value of the dichotomous variable $St+$ - but I think this might be clearer to the reader if the equation is written using definition by cases notation (<https://planetmath.org/definitionbycases>) rather than the implicit multiplication by zero trick (of course they give equivalent results, but this way reinforces to the reader that $St+$ is dichotomous).

We agree that this makes this equation easier to read. For consistency, we have also updated equation 10, 13 and 15.

- Line 298-333: The discussion of the defined latent state is important but perhaps out of place in the materials and methods section?

We agree. This also duplicates ideas that are written in the discussion. These 2 paragraphs have been removed.

- Lines 369-370: What about values between 35 and 60?

The rules used in the programme are actually complicated with 27 possible combinations of 3 categories (low / medium / high antibody levels) over 3 consecutive testing campaigns. We did not want to go into too much detail about these as it could be confusing and would bring nothing to the content of the paper.

- Lines 384-388: Technically it is not being "Bayesian" that takes the time, but relying on computationally intensive fitting methods (i.e. MCMC). Also, could you briefly mention the more Bayesian methods of approaching this e.g. stochastic variable selection - see "O'Hara, R.B., Sillanpää, M.J., 2009. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.* 4, 85–117. <https://doi.org/10.1214/09-BA403>." I don't think these would be computationally feasible in your case, but it is still a potential alternative to resorting to frequentist methods of fitting a simplified version of the model.

This has been added towards the end of the discussion: *The identification of the most predictive time interval between risk factor occurrence and seroconversion, required the evaluation of the associations between the probability of seroconversion on a given month and risk factor occurrence over all possible intervals between this month and the 24 previous months. Although there are several Bayesian methods for such variable selection O'Hara (2009, estimation using MCMC is time consuming and was not feasible in our case.* (See L759 of the current version)

- Line 401-402: But as I understand you are using a log link function within the GLM, so by including an additive effect of $\log(N)$ on the log scale you still end up with a multiplicative effect of N on the scale of the response distribution, so I am not entirely sure that this "allows a decreasing effect of each animal" as you state. If you wanted to do this perhaps you could include a 2nd degree polynomial of $\log(N)$? Or have I misunderstood something?

This was to attempt at explaining why the log of the variable was providing a better fit than the untransformed variable. What was meant was that the increase is steeper towards the beginning of the curve. This effect is visible and plotted at the bottom of Figure 7.

- Line 640: Perhaps "failure" rather than "absence" of convergence?

Changed. (L642)

- Lines 643-647: Would multiple imputation (i.e. running the MCMC sampler 20 times using 20 different fixed values of Se/Sp from across the prior) be an alternative to allow these models to be fit? Or perhaps using a different sampling strategy that is more efficient in these situations such as Hamiltonian Monte Carlo? Obviously the latter is not possible in JAGS, but multiple imputation has been done in these situations.

The Stan implementation, which uses Hamiltonian Monte-Carlo, has resolved this by allowing wider priors on test characteristics.

- Line 720: Please do not use the word "uninformative" in relation to priors - there is no such thing. Either "minimally informative" or "non-informative" is fine.

Changed

- Could you give some JAGS code for (possibly simplified versions of) models 1-4 as an appendix? I realise that this is contained within the GitHub repo, but as far as I can see an interested reader needs to mentally paste together the R code that generates the JAGS model to be able to assess the code. It would be easier if the model representation was available directly.

We have now provided the JAGS and Stan code for the model in S1. Also, we have added a `write_model` argument to the model call that writes a text file containing the model code.

Finally, I would like to apologise for the lengthy delay in providing this review, which was due to an unexpectedly increased workload at my end - sorry.

No problem. Thank you for the useful comments. It took us even more time to reply.