WILDLIFE HEALTH ECOLOGY RESEARCH ORGANIZATION
San Diego, United States
www.wildlifehero.org
bennyborremans@wildlifehero.org

Response to reviewers - PCI Ecology
8 June 2024
Reconstructing prevalence dynamics of wildlife pathogens from pooled and individual samples
[Revision 2]

Dear Recommender and Reviewers,

We would like to thank the Recommender and both Reviewers for taking the time to review our revised manuscript. We have addressed the one remaining comment and are happy to share that we found a great solution to the double likelihood problem, please see below for details.

Sincerely,
on behalf of all authors,
Benny Borremans

WILDLIFE HEALTH ECOLOGY RESEARCH ORGANIZATION
San Diego, United States
www.wildlifehero.org
bennyborremans@wildlifehero.org

Recommender:
Both reviewers recognize the improvements made to the manuscript after this round of revision. One of the reviewers is concerned about the fact that changes to eq. 1-2 requires additional simulations and discussion; after reading their review, I tend to share this concern, and I am expecting the revised version to address these comments in detail.

Reviewer 1:
The authors have successfully addressed all of my comments and I believe that the manuscript is improved in both content and readability. I am happy for the manuscript to be accepted in this state.

Reviewer 2:
The changes the authors make have greatly improved the manuscript. The extra information and clarification they provide address all of my comments. However, the additional information they provide about equations 1 and 2 make it clear that the model formulation must be remedied before publication. This will require rerunning simulations, etc. since the model will need to be refit following reformulation. As it is written, unfortunately the model is not mathematically well defined, leading to potentially incoherent situations. For example, if you want to simulate data for a population in which prevalence is equal to 0, but simultaneously all members of the population have a covariate value that is known to perfectly correlate with disease, then you have a problem. The prevalence parameter implies all y_i should equal 0, while the covariate implies all y_i should be equal to 1. The model needs to avoid such mathematically pathological situations, and it is unreasonable to simply advise against use in such scenarios in a discussion section for more substantial reasons described below.
Standard probability rules require a random variable to have a single definition. Defining it twice, as the authors do via equations 1 and 2 is not a valid way to specify a joint probability model. The authors may only specify a single Bernoulli distribution for y_i. The authors must explicitly write mathematically, in an equation, how they wish the individual-level covariates to interact or otherwise relate to population-level prevalence. Although it is simple to write a single equation involving both theta and the covariates, the revision could potentially place the population-level prevalence term in tension with the effect of individual level covariates. Naturally, population-level prevalence partly arises from the aggregate effects of individual-level risks and outcomes. A revision will likely require some careful thinking about how to (re?)interpret model parameters.
Statistical software may or may not enforce the one definition rule, but that does not justify ignoring it. For example, while the R software Stan may interpret multiple definitions for a random variable as a user's request to make two separate contributions for the variable to the log-likelihood, the R software Nimble refuses to build models when variables are defined twice. In general, statistical software is not necessarily provided with guard rails to prevent users from doing "prohibited" things. Drawing on an example commonly taught in introductory regression modeling classes, the R function lm() (and equivalent functions in other statistical software, like SASS) will let users fit a least squares linear regression model to binary data, even though users should use logistic regression models (or similar) to fit binary data. When software provides estimates for models that are not mathematically well defined, the estimates

WILDLIFE HEALTH ECOLOGY RESEARCH ORGANIZATION
San Diego, United States
www.wildlifehero.org
bennyborremans@wildlifehero.org

in general will not be able to be interpreted in the way users intend. Software predictions may also be non-sensible, which is the main concern with fitting binary data to a least square linear regression.

Author reply:

This comment is regarding the following 2 equations:

$$y_i \sim Bernoulli((1-\psi) * \theta_{t[i]}) \quad (1)$$

$$y_i \sim Bernoulli((1-\psi) * logit^{-1}(\beta_0 + \sum_{z=1}^{k}(\beta_z * x_{z,i}))) \quad (2)$$

Where y_i is the shedding status (0 or 1) of individual *i*, ψ is the false negative rate correction parameter, $\theta_{t[i]}$ is population prevalence at time *t*, β_0 is the regression intercept, β_z is the coefficient for covariate *z*.

We understand and agree with the reviewer's concern about this double likelihood, and would like to thank the reviewer for pointing it out so clearly.
We are happy to share that we found a solution for this non-trivial problem, that absolutely improved the model.

The only definition of outcome variable y_i is now the regression model, eq. (2) above (or equation 1 in the manuscript), adapted slightly so that intercept β_0 is time-specific:

$$y_i \sim Bernoulli((1-\psi) * logit^{-1}(\beta_{0,t[i]} + \sum_{z=1}^{k}(\beta_z * x_{z,i})))$$

To then calculate the mean probability of success for each sampling time, we use Monte Carlo integration over all covariate values. This is a numerical approximation for full integration, which would be feasible for a single covariate, but becomes exponentially more computationally intensive with more covariates. Monte Carlo integration randomly generates covariate samples from their respective distributions and calculates success probability for each random combination of samples. The mean of these success probabilities is the prevalence estimate:

$$\theta_t^{ind} = \int_X logit^{-1}(\beta_{0,t[i]} + \sum_{z=1}^{k}(\beta_z * x_{z,i}))p(X)dX \approx \frac{1}{z}\sum_{j=1}^{z} logit^{-1}(\beta_{0,t[i]} + \sum_{z=1}^{k}(\beta_z * x_{z,i}))$$

The distributions of each covariate can be any type. For the simulation model we used a normal distribution, with distribution parameters estimated as part of the model. Finally, θ_t^{ind} is used to inform overall prevalence through a straightforward Beta distribution:

$$\theta_t \sim Beta(\theta_t^{ind}\kappa, (1-\theta_t^{ind})\kappa)$$

As you can see in the text, this improved approach avoids the double likelihood, is able to recover all parameters well, and shows great results.

WILDLIFE HEALTH ECOLOGY RESEARCH ORGANIZATION
San Diego, United States
www.wildlifehero.org
bennyborremans@wildlifehero.org

Please see page 6 and 7 (and eq. 6 on page 12) for these updates.
We also updated the model schematic illustration, re-ran all models and updated the results and figures. All conclusions remained the same.