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

Response to reviewers - PCI Ecology
20 Feb 2024
Reconstructing prevalence dynamics of wildlife pathogens from pooled and individual samples
[Revision 1]

Dear Recommender and Reviewers,

We would like to thank the Recommender and both Reviewers for taking the time to thoroughly review our manuscript and provide thoughtful comments.
All comments were addressed and implemented, resulting in an improved version of the manuscript we are proud to resubmit.

Please see below for detailed replies to all comments.

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 are generally very supportive of the work, and make specific suggestions to improve the presentation. After reading the preprint and the reviewer comments, I find myself in agreement with their feedback. It should be possible for the authors to incorporate all of it in a revision.

Reviewer 1:
This study describes a new computational approach for obtaining viral prevalence estimates using naturally pooled samples, the use of which is currently limited to presence/absence information in an area.
The methods presented represent a significant advance in the analysis of pooled samples and have the potential to allow the study of pathogen prevalence in wildlife populations without the time, expense, and hazard of catching and handling individual animals. Where individual level samples are also available, the authors include methods for directly incorporating individual covariates. The true prevalence over time section of the model offers the potential to, although not explored by the authors, directly incorporate a transmission model and fit these parameters which should allow for the direct combining of uncertainty from the other sections of the model. The discussion provides a fair assessment of the potential utility of the model whilst discussing its limitations, specific requirements, and drawbacks.

R1.1. Whilst the authors explore a good range of non-ideal data scenarios, I am not convinced that these scenarios reflect a 'realistic' dataset. I would like to see if the close match to the simulated prevalence is maintained when multiple potentially confounding factors are combined, i.e., small and varying sample sizes, taken at irregular intervals. I would recommend not overstating the realism of the test data in the discussion. Below are some suggestions for clarifications of the text and figures.
Author reply. Thank you for pointing this out. We updated the text so it explicitly states that the prevalence fluctuations are not based on any particular system, but rather were chosen so that a range of sample availability scenarios can be tested. We do believe though that the sample availability scenarios are realistic, with regards to timing and sample size, and is based on our direct experience with a surveillance dataset on Hendra virus in flying foxes in Australia.
We agree that adding scenarios that combine multiple factors would provide useful additional insights, so we added a scenario in which sampling sessions are irregular, asynchronous (pooled vs individual) and with lower sample sizes per session, and a scenario in which sampling sessions are irregular, asynchronous and with fewer sampling sessions. We found that the model can handle these scenarios and performs as well as can be expected given the data availability. We added these scenarios to the supplementary information and results, and added reference to these results to the discussion.

R1.2. I'm not sure of the length allowances for the abstract in this journal, but I find the abstract rather long, which detracts rather than enhances interest in the article.
Author reply. We made the abstract more concise.

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

R1.3. The introduction describes in detail the current state of the field, and the need for the model. The research question is clearly presented but could again be more concise for readability.
Author reply. Thank you for this suggestion, we agree and were able to shorten this section substantially.

R1.4. Figure 1 – The black lines on the figure showing the connectivity between the sections aren't very informative and I think make things less clear. That the relevant parameter is highlighted in a different colour is enough to see that it occurs in all three sections. Perhaps make it bold as well if the colour alone is not clear enough.
Author reply. Good point, we removed the black lines.

R1.5. Is θ observed as per the key? Is that not what all sections of the model come together to estimate? If the important thing is that it's estimated by all of them, put it outside of the other boxes.
Author reply. That indeed works better, thank you for the suggestion.

R1.6. Should equation 2 match the relevant equation in figure 1 (currently the yellow equation in box C)?
Author reply. Those equations should indeed match, we are happy you spotted this error and updated the equation in the figure.

R1.7. Page 10 line 5 – the three key factors that influence final pooled concentration. You mention a few paragraphs later than urine volume is assumed to be equal, but on first read-through I was wondering in this section why it was being ignored. A simple line of 'here we focus on these first two factors' would do to make readers stop wondering where the third one was!
Author reply. Good suggestion, we added this.

R1.8. Where laboratory experiments are required to determine distributions/baseline values should be clarified.
Author reply. Laboratory experiments are not required for any aspect of the method. The distribution of Ct values that is used for constructing the Ct probability function is typically determined using samples from wild individuals. We mention this in the relevant section on page 11: "Individual samples, if collected, can be used to inform this distribution." For clarity though we added "wild" to the relevant mention of this in the discussion on page 24.

R1.9. Add to the discussion on pool size limitations of this method that this should be taken into account during the field experimental design and set-up if feasible for most reliable results.
Author reply. We added this to the relevant paragraph on page 11.

R1.10. Possibly mention that this is more suitable for some wildlife species than others given their usual behaviours and living arrangements.

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

Author reply. We added this to the relevant methods section on page 19 ("A sampling design that incorporates pooled samples will be more beneficial for some wildlife species than for others, but there are no inherent limitations to which species this approach could be applied to."),

R1.11. Figure 2 – Would add either in the figure or the legend, how these steps (A-D) relate to the numbered steps in the main text. Am I understanding correctly that example in this figure (Ct 36, 2/3 bats/ 20% prevalence) is randomly chosen, and that this process would need to be completed for all possible combinations? If so, add a final line in the legend stating this?
Author reply. We originally did exactly that, but relating the steps in the text to specific parts of the steps in the illustration turned out to be quite confusing. For this reason we outlined the specific steps in the figure legend. This process is indeed repeated for all possible combinations, and we added this to the legend as suggested.

R1.12. Figure 4 – I think that the 50% CI shading should stand out a bit more. I would also recommend moving the datapoints to the top layer of the figure so that they are not hidden behind the fitted prevalence curves or the credible interval band.
Author reply. Those modifications indeed make the figures better, thank you. We updated the relevant figures in the main text as well as in Supplementary Information.

Reviewer 2:
The authors propose a method to estimate prevalence of an infectious disease using pooled samples. The method can accommodate time-varying prevalence as well as covariate effects. The authors introduce and review previous work on the subject, noting that the "dilution effect" is the central challenge for accurately estimating prevalence from pooled samples (or a combination of pooled and individual samples). The authors frame prevalence estimation as a hierarchical modeling problem. A Gaussian process (GP) and regression coefficients model time-varying prevalence trends and covariate-driven deviations. A mixture model accounts for the effect of pooling on sample pathogen measurements. The authors use simulation to demonstrate the method and provide some discussion of sensitivity to model mis-specification; the method is not applied to real data.
Overall, the manuscript's technical content is presented clearly but its contribution is unclear due to strong methodological similarities with literature cited in the introduction.
Major comments:
R2.1. How do the submitted manuscript's contents make contributions beyond previous literature cited in the introduction? The authors suggest in their abstract and introduction that a key limitation of existing work is an inability to "...estimate the proportion of positive individuals using concentration when the underlying distribution of test values does not follow a standard-family (e.g., Gaussian) distribution..." The authors also suggest existing work is not able to accommodate data from both individual and pooled samples simultaneously. The authors suggest Cleary et al. (2021) and Self et al. (2022) are the closest comparison methods, which still lack features the submitted manuscript proposes. In particular, Self et al. (2022) discusses

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

adapting much older methods proposed in Zenios and Wein (1998) and companion paper Wein and Zenios (1996).

Author reply. We would like to thank the reviewer for their close reading, which prompted us to revisit the literature and reflect on what we believe makes our work distinctive.

We agree with these comments, which made us realize that we had not been sufficiently precise in communicating the novel aspects of the model. The model introduces two main advances: (1) the complete numerical calculation of the probability density function used to estimate prevalence from the concentration of pooled samples, taking into account all possible combinations of negative and positive individuals, and the underlying distribution of concentrations in the population; (2) the use of the combination of pooled and individual samples to estimate false negative rate, which is then used to account for the biased estimation of regression coefficients that occurs due to the inclusion of false negatives. Additionally, and perhaps most importantly, our manuscript is written for an audience of researchers working with wildlife, with the intent of bridging the gap between hard statistics and applications to wildlife systems.

We edited all text to better reflect this.

With respect to Zenios and Wein(1998)

- **R2.2.** How does use of equation 5 in the submitted manuscript differ from equation 2 in Zenios and Wein? Equation 5 in the submitted manuscript appears to present one of the authors' main contributions. However, equation 5 appears to be identical to equation 2 in Zenios and Wein, despite changes in variable names and some notation.
  Author reply. The equations are indeed identical in essence. We make no claims however about this equation being novel, so for clarity we added a reference to the Zenios and Wein equation.

- **R2.3.** How do the submitted manuscript's distributional assumptions differ from those in Zenios and Wein? The submitted manuscript claims to be unique in proposing analytic methods that can support non-standard distributions for individual-level concentrations (pg. 11), but Zenios and Wein already appear to handle such cases (Section 2). Zenios and Wein work in a general probabilistic framework that does not restrict their formulation to Gaussian distributions or other parametric families.
  Author reply. We agree that Zenios and Wein do not restrict their formulation when proposing the general approach, so we adapted the text to better reflect this.

- **R2.4.** How do the submitted manuscript's computational methods differ from those in Zenios and Wein? Zenios and Wein propose using Monte Carlo methods to facilitate computation for non-standard distributions for individual-level concentrations (Sections 4, 7.3). Similarly, the submitted manuscript recommends Monte Carlo methods to evaluate equation 5 (pg. 12) when the exhaustive computation discussed first is computationally infeasible (pg. 10). Zenios and Wein do discuss using the central limit theorem to motivate computationally faster Gaussian approximations when pooled samples contain material from many individuals, but this does not appear to be a

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

required computational technique or distributional limitation of the earlier work (Section 4.1).
Author reply. Zenios and Wein, as well as Self et al, indeed propose computational methods that allow estimation of the probability densities of concentrations. The main difference with our approach is that we introduce an algorithm that does not rely on estimation, but rather considers every possible combination of negative and positive individuals while taking into account the underlying distribution of concentrations in the population. We edited the text to better reflect this difference, in particular now highlighting that our method complements the existing ones.

With respect to Cleary et al.(2021)
Author reply. We believe that the reviewer meant to reference Self. et al 2022, and not Cleary et al. 2021, based on the comments (and cited page number) below.

- **R2.5.** How does use of equation 5 in the submitted manuscript differ from the mixture distribution presented near the top of pg. 4685 in Cleary et al.? Similar, in some regards to comments regarding Zenios and Wein, Cleary et al. uses the distribution to model pooled sample concentrations while mixing over 1) the unknown number of true positive samples in the pool without 2) making Gaussian assumptions about the individual-level concentration distributions (referred to as “biomarker concentration” distributions in Cleary et al.). Cleary et al. only make an assumption that the observed, pooled concentration value is observed with Gaussian measurement error—the underlying distribution for the true concentration is a mixture over biomarker concentrations that appears to be identical in spirit if not content to equation 5 in the submitted manuscript.
  Author reply. We agree that the equations are quite similar, but they do differ in that Self et al use a mixture model approach (that is better suited for dealing with the negative/positive cutoffs that are particularly vague for antibody concentrations), whereas our approach does not.

- **R2.6.** Is the submitted manuscript novel in its use of pooled and individual samples? The framework Cleary et al. proposes to model pooling does not appear to limit application of the methods to individual samples. For example, their method seems to allow an individual sample to be represented as a “pool of size 1” with no measurement error.
  Author reply. Our approach is indeed not novel in combining pooled and individual samples, and we don’t make that claim either. We believe our approach is novel in how it uses that combination of pooled and individual data to estimate false negative rate, which in turn enables accounting for the biasing effect of false negatives on the estimation of coefficients in the individual-level regression model.

- **R2.7.** The second to last introductory paragraph in the submitted manuscript seems to imply its use of generalized linear modeling structures to include individual-level covariates, and (basic) Gaussian processes to non- parametrically model time-varying processes is novel. However, equation 1 in Cleary et al. also specifies a generalized linear

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

model structure for individual-level infection, which could reasonably include individual-level covariates and non-parametric components (i.e., such as splines) that could potentially model time-varying prevalence with similar flexibility as basic Gaussian processes.
Author reply. We didn't intend to imply that these model aspects are novel, so we edited the text to better reflect the advances introduced by the model: "The model offers two key advances: first, the ability to estimate the false negative rate ensures that the effect coefficients of infection covariates can be estimated correctly, as these can otherwise be strongly affected by the presence of false negative samples. The second is the introduction of an algorithm that enables the full numerical calculation of the probability density function of concentrations in pooled samples."

R2.8. The submitted manuscript discusses on p.11 how the weighting function P(C_j|...) in Equation 5 is rarely uniform and often unknown in practice. In simulation, the authors briefly discuss how estimates are biased when the weighting function is misspecified (p.20). Can additional details be provided, alongside methods or recommendations for how to estimate the weighting function or find estimates for it in existing literature? The weighting function seems like a critical component for the proposed method's success, with relatively little concrete guidance or demonstration for how to work with it in application to real datasets.
Author reply. Thank you for this suggestion, we now included a warning and specific guidance on the use of this weighting function.

Minor comments:
R2.9. The submitted manuscript points out that concentration data can be more informative than positive/negative test result summaries of concentration data, especially for pooled samples. The submitted manuscript essentially uses concentration data to mitigate the risk of false positive/negative outcomes if concentration data were simplified to binary outcomes. But, would individual-level samples not also benefit from modeling concentration data instead of binary outcomes? Can some discussion be added to help explain or justify the motivation to only use positive/negative summaries of concentration data for individual-level samples?
Author reply. Thank you for this suggestion. The concentration of individual samples could indeed be used to include an observation process layer to the model that takes into account the variation that can be introduced when using the concentration to classify a sample into negative or positive. We now added this to the text: "Note that while a Ct value is generated for individuals contributing to a pooled sample, the individuals used for the "individual sample" model described in the previous paragraph only have a negative or positive status, and not a Ct value. When required it is possible to add an observation process layer to the model that explicitly models the classification of sample into negatives or positives based on the concentration, as for example shown in …"

R2.10. Equations 1 and 2 both define the response variable y_i, which seems redundant and potentially confusing to readers. Equation 2 basically appears to repeat Equation 1, but by with one definition for \theta expanded. Can Equation 2 be simplified to explicitly define \theta instead of redefining y_i? If not, is some other presentation possible?

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

Author reply. We understand the suggestion, but have a good reason for using this formulation, which is that theta is a parameter that is shared (see Figure 1), and can therefore not be equated to the corresponding component of the individual-level regression model. We added a brief explanation of this to the relevant methods section.

R2.11. It is convention in statistics to write the true/unknown/theoretical value of model parameters without "hats", and estimates of model parameters with hats. The submitted manuscript appears to break from this convention. For example, almost all key model parameters are exclusively written with hats, regardless of context, such as \hat\theta, \hat\varphi, \hat\beta. Is it possible to review the use of hat notation?
Author reply. You are correct, so we removed the hats.

R2.12. On page 14 after equation 6, the mean function for the Gaussian process W_t is written as a bold number 0. It seems like this is a typo, since bold symbols are conventionally reserved for vectors, but the specified GP mean function is univariate here.
Author reply. The Gaussian Process is modeled as a zero-mean process, with the zeros being a vector, hence the bold font.

R2.13. On page 14 in the paragraph starting "A useful property..." the authors evaluate the Gaussian process covariance function (eq. 6) and incorrectly label the evaluated function as correlation.
Author reply. Corrected.

R2.14. On pages 14 and 15 the authors discuss choosing simulation parameters that "would result in realistic prevalence fluctuations." Can the authors provide specific citations or references to clarify which disease systems or outbreaks helped inspire the simulation's prevalence curves?
Author reply. While the sampling design was based on a specific system (bat-borne viruses, as referenced in the text), the prevalence fluctuations were not, and were instead chosen to provide a useful range of prevalence fluctuations to test model performance. We changed the text so that this is explained better.

R2.15. Similarly, the authors state they run the simulation for "a time period of 300 (an arbitrary number) time points...." It would be helpful if the authors could be more precise. For example, do the time points nominally represent seconds, hours, minutes, days, weeks, or months? Specifying a time scale for the simulation should be able to help 1) make the simulation more convincingly realistic and 2) better clarify the types of applications and sampling requirements the method is being developed for.
Author reply. The time units are arbitrary, and it does not matter for the model what the units are. We edited the text to explain this better.

R2.16. The manuscript text at the top of page 19 describes Figure 4A as presenting 95% credible intervals for prevalence, but the caption for Figure 4A says it shows 50% credible intervals. Can the typo or figure be remedied?

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

Author reply. Corrected, thank you for catching this.