Review for PCIEcology #598: "Reconstructing prevalence dynamics of wildlife pathogens from pooled and individual samples"

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:

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

- 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 required computational technique or distributional limitation of the earlier work (Section 4.1).

- With respect to Cleary et al. (2021)

  - 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.

  - 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.

  - 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 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.

- 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.

Minor comments:

- 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?

- Equations 1 and 2 both define the response variable y_i, which seems redundant and potentially confusing to readers. Equation 2 basically appears to repeats 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?

- 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?

- 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.

- On page 14 in the paragraph starting "A useful property…" the authors evaluate the Gaussian process <u>covariance</u> function (eq. 6) and incorrectly label the evaluated function as <u>correlation</u>.

- 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? 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.

- 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?

References

Cleary, B. et al. Using viral load and epidemic dynamics to optimize pooled testing in resource-constrained settings. Science Translational Medicine 13, eabf1568 (2021).

Self, S., McMahan, C. & Mokalled, S. Capturing the pool dilution e↵ect in group testing regression: A Bayesian approach. Statistics in Medicine 41, 4682– 4696 (2022).

L. M. Wein, S. A. Zenios, Pooled testing for HIV screening: Capturing the dilution effect. Oper. Res. 44, 543–569 (1996).

S. A. Zenios, L. M. Wein, Pooled testing for HIV prevalence estimation: Exploiting the dilution effect. Stat. Med. 17, 1447–1467 (1998).