Reconstructing prevalence dynamics of wildlife pathogens from pooled and individual samples

Benny Borremans[1,2,3,*], Caylee A. Falvo[2], Daniel E. Crowley[2], Andrew Hoegh[4], James O. Lloyd-Smith[5], Alison J. Peel[6], Olivier Restif[7], Manuel Ruiz-Aravena[2,6,8], and Raina K. Plowright[2]

[1]Wildlife Health Ecology Research Organization, San Diego, United States
[2]Department of Public and Ecosystem Health, College of Veterinary Medicine, Cornell University, Ithaca, United States
[3]Evolutionary Ecology Group, University of Antwerp, Antwerp, Belgium
[4]Department of Mathematical Sciences, Montana State University, Bozeman, USA
[5]Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, United States
[6]Centre for Planetary Health and Food Security, School of Environment and Science, Griffith University, Queensland, Australia
[7]Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom
[8]Department of Wildlife, Fisheries and Aquaculture, Mississippi State University, Mississippi State, United States
[*]bennyborremans@gmail.com

Last update: February 19, 2024.

Abstract

Pathogen transmission studies require sample collection over extended periods, which can be challenging and costly, especially in the case of wildlife. A useful strategy can be to collect pooled samples, but this presents challenges when the goal is to estimate ~~infection prevalencedynamics. In particular, pooling typically results in~~ prevalence. This is because pooling can introduce a dilution effect where ~~mixing positive and~~ pathogen concentration is lowered by the inclusion of negative or lower-concentration samples~~lowers the overall concentration. Simultaneously,~~ , while at the same time a pooled sample ~~is more likely to test positive as this requires only one or a few positives. The~~ can test positive even when some of the contributing samples are negative. If these biases are taken into account, the concentration of a pooled sample can be ~~used~~ leveraged to infer the most likely proportion of positive individuals, and thus improve overall prevalence reconstruction, but few methods exist that account for the sample mixing process~~and none can handle common but non-standard frequency distributions of concentrations~~.

We present a Bayesian multilevel model that estimates prevalence dynamics over time using pooled and individual samples in a wildlife setting. The model explicitly accounts for the complete mixing process that determines pooled sample concentration, thus enabling accurate prevalence estimation even from pooled samples only. As it is ~~nearly impossible~~ challenging to link individual-level metrics such as age, sex, or immune markers to infection status when using pooled samples, the model also allows the incorporation of individual-level samples. ~~These are used to further improve prevalence estimates and estimate variable correlations.~~ Crucially, when individual samples can test false negative, a potentially strong bias is introduced that results in ~~wrong regression coefficient estimates~~ incorrect estimates of regression coefficients. The model, however, can ~~use~~ account for this by leveraging the combination of pooled and individual samples~~to estimate false negative rate and account for it so that regression coefficients are estimated correctly~~. Last, the model enables estimation of extrinsic environmental effects on prevalence dynamics.

Using a simulated dataset ~~based on~~ inspired by virus transmission in flying foxes, we show that the model is able to accurately estimate prevalence dynamics, false negative rate, and covariate effects. ~~Using~~ We test model performance for a range of ~~scenarios based on real study systems we show that the model is highly robust~~realistic sampling scenarios and find that while it is generally robust, there are a number of factors that should be considered in order to maximize performance.

The model presents an important advance in the use of pooled samples for estimating prevalence dynamics in a wildlife setting, can be used with any biomarker of infection (Ct values, antibody levels, other infection biomarkers) and can be applied

to a wide range ~~of human and wildlife pathogen study~~ ofhost-pathogen systems.

Introduction

When monitoring and studying pathogen transmission over time, it is essential to estimate prevalence dynamics. Prevalence, defined as the proportion of individuals in a population that tests positive for the current (e.g., presence of a pathogen or its genetic material) or past (e.g., antibody presence) presence of an infectious organism, is a key metric, yet can be difficult to estimate. The reason for this is that it is almost never feasible to test every individual in population, which means prevalence needs to be estimated from a population subset. As a result, methods are needed to estimate prevalence from imperfect data due to constraints in the number and quality of samples.

Sampling will depend on constraints (logistical, technical, individual availability, monetary), and different sampling strategies can be used to maximize the number of individuals being sampled [**RN17**, **RN20**]. One such strategy is to pool samples, either by combining samples collected from different individuals (which reduces resource investments in testing and collection; [**RN18**]), or by collecting samples that already consist of material from multiple individuals (e.g., monitoring of SARS-CoV-2 in sewage; [**RN19**]). In studies of wildlife disease this latter approach is relatively common, for example when collecting fecal droppings in a den or cage containing multiple animals [**RN20**], or when collecting water samples in a lake or in wastewater [**RN21**]. An important drawback of the latter approach to pooling is that the sample cannot be linked to individual-level data, except indirectly under certain controlled conditions; [**RN94**].

Individual samples provide the highest-resolution information, as they allow additional individual-level data to be collected, including body measurements, estimates of sex and age class, and a wide range of biomarkers such as antibodies, blood proteins or other infections. These additional data are highly valuable as they can be used to learn more about correlates and drivers of infection. Depending on the study system, however, there can be several challenges to collecting and interpreting individual samples. A first is that the collection and processing of individual samples can be costly — in terms of effort, time or monetary resources — which limits sample sizes and temporal/spatial resolution. It can also be difficult to capture and sample individuals, for example when dealing with species that are elusive or live in low-density populations. Another challenge can arise when individuals do not shed a pathogen continuously but intermittently because of fluctuating pathogen concentrations. For example, the rodent *Mastomys natalensis* is known to shed arenavirus in varying concentrations [**RN29**]. Intermittent shedding means that it is possible to collect a negative sample or a sample with an undetectable pathogen concentration even though the individual can be considered infectious, leading to false negative results with regards to determining whether or not an individual is

infectious.

A powerful study approach is to optimize the trade-off between sampling cost and data resolution by collecting both pooled and individual-level samples. This is commonly done in bat pathogen studies, where samples are collected from individual bats using net captures — which enables the collection of high-quality samples and associated individual variables — as well as from multiple bats simultaneously using plastic sheets placed under roosts [**RN22**, **RN23**, **RN24**]. This approach is particularly useful when the goal is to estimate prevalence dynamics.

When estimating prevalence, the use of pooled samples presents two well-known challenges, both resulting from the fact that multiple individuals contribute to the same sample. The first challenge is that a pooled sample can test positive regardless of how many of the contributing individual are actually positive. As a result, the proportion of positive pooled samples can be biased upwards, leading to overestimates of prevalence [**RN24**]. The second challenge is the opposite of the first, and is the fact that a pooled sample can test negative even when one or multiple contributing individuals are positive. This can arise when the sample is diluted by negative samples, causing the concentration of the positive sample(s) to lower and fall below a detection threshold (which is called the dilution effect in pooled/group/-composite testing literature; [**RN11**]). Assay sensitivity will be an essential factor in how low the diluted concentration can be before it can no longer be detected. Several approaches have been suggested to deal with these two challenges [**RN11**, **RN8**], the most recent of which presents a Bayesian mixture model approach that can account for both at the same time under certain conditions [**RN10**]. Most studies on the analysis of pooled samples focus on testing protocols for cost reduction, with the goal of eventually identifying the positive individuals [**RN18**, **RN25**, **RN26**]. Perhaps for this reason, few methods have been developed for explicitly using pooled samples to estimate prevalence in the population [**RN5**, **RN6**, **RN8**, **RN13**, **RN24**], and even fewer have attempted to use the actual concentration of the infectious agent (or another biomarker like antibody concentration) in the pooled sample to estimate how many of the contributing individuals are positive ~~. Furthermore, to our knowledge no methods exist that are able to estimate the proportion of positive individuals using concentration~~ [**RN10**, **RN8**, **RN9**]. A particular challenge arises when the underlying distribution of test values does not follow a standard-family (e.g. Gaussian) distribution, even though this is the most common situation, especially for wildlife populations [**RN39**, **RN40**]. Few methods exist that can incorporate such distributions, and to our knowledge none provide a method for numerically calculating the full probability distribution of test values, instead using approximation methods [**RN9**, **RN10**]. Leveraging the information present in the concentration of the infectious agent in pooled samples instead of only using binary negative/positive information ~~, however, could~~ can lead to significant

improvements in the estimation of prevalence, particularly in the case of disease surveillance in wildlife populations.

We present a multilevel Bayesian modeling approach to estimate infection prevalence simultaneously from both individual and pooled samples, explicitly using the concentration of the infectious agent in pooled samples and thereby accounting for the biological mixing process that generates pooled sample concentrations. ~~Prevalence is modeled over time, and the model leverages the fact that prevalence values closer in time are likely to be more correlated than those separated by longer times. The multilevel model further includes regression models at both the individual level and the temporal/population level, thus allowing the inclusion of covariates such as antibody levels, age or sex that can provide additional information about prevalence while testing whether they correlate with individual-level infection status or with larger-scale ecological processes that drive prevalence fluctuations. Last, because of the use of both pooled and individual-level data, the model is able to provide estimates of false negative rates in individual-level samples, which can occur due to factors such as intermittent shedding, low/undetectable concentration, sampling conditions or assay characteristics. Through the incorporation of covariates the model also offers the opportunity to predict the false negative status of specific individual samples. In summary, the model offers~~ The model presents 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 ~~that by explicitly modeling the biological mixing process that generates the concentration values of pooled samples, it becomes possible to estimate prevalence even in the absence of individual-level data~~the introduction of an algorithm that enables the full numerical calculation of the probability density function of concentrations in pooled samples.

Model use and performance is presented using simulated data ~~based on~~ inspired by a bat-pathogen study system, but we highlight that this approach can be used for any situation in which prevalence fluctuations are estimated from pooled samples with a known (or estimated) number of contributing individuals, especially when combined with individual samples. To illustrate the broader relevance, and test how the model performs under different conditions, we included relevant scenarios that each resemble a realistic biological situation. The approach presented here is particularly useful when the goal is not to identify which specific individuals are positive but to determine prevalence in the population, because there is no need to re-test de-pooled samples. Examples include monitoring SARS-CoV-2 prevalence [**RN13**, **RN8**], estimating prevalence in wastewater if the number of contributing individuals can be estimated [**RN21**], assessing pathogen prevalence in the animal production industry [**RN27**], or estimating pathogen prevalence in wildlife populations [**RN28**]. Note

that while the example presented here focuses on infection prevalence, the model can also be applied to other biomarkers such as antibodies.

Methods

The main goal of this study is to estimate the true, unknown, proportion of pathogen-positive individuals over time, from both pooled and individual samples. Each of these types of samples presents a challenge for estimating prevalence, but also an opportunity, as outlined in Table **??**. Note that the focus is on "naturally" pooled samples, where collection was not done directly from individuals, as opposed to "technically" pooled samples that were pooled intentionally after collection from individuals.

Table 1: Sample types and their different challenges and advantages for estimating prevalence.

Sample type	Challenge	Advantage	Example
Pooled.	Number of positives unknown; Number of contributors possibly unknown; Dilution effect.	Sample multiple individuals at once. Lower collection/testing cost per individual.	Blood sample pooling to reduce testing costs; Urine collected from sheets under a bat roost.
Individual.	False negatives possible.	Additional individual measurements	Samples collected in combination with individual data such as sex, age and body weight; Urine, blood samples and body measurements collected from captured bats.

Here, we simulated data ~~based on~~ inspired by existing studies on flying foxes for research on temporal virus dynamics [**RN22**, **RN23**, **RN24**]. For the reasons mentioned above, bat virus studies often use field sampling designs that rely heavily on the collection of pooled urine and fecal samples under bat roosts [**RN23**]. 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. We chose to use simulated data only, as the goal of this study is to present and test a model to estimate prevalence, which can be done optimally when all underlying parameters are known and different scenarios can be generated. This makes it possible to determine how well the model is able to estimate the known parameters and prevalence dynamics for a range of scenarios. The simulated datasets are described below at the end of the Methods section.

The model is described in three parts, representing the multilevel/hierarchical nature of the model (**??**). The two main parts, a model for estimating prevalence from individual samples and a model for estimating prevalence from pooled samples, are linked by a third model of true, unobserved prevalence dynamics. We used a Bayesian multilevel model (also called a hierarchical model), as this provides a solid framework for linking the different model components, modeling unobserved latent parameters, incorporating prior knowledge through prior distributions, and providing posterior distributions of parameter estimates that show the uncertainty. While not done here, it would be straightforward to include an additional observation model that takes into account observation/measurement errors.

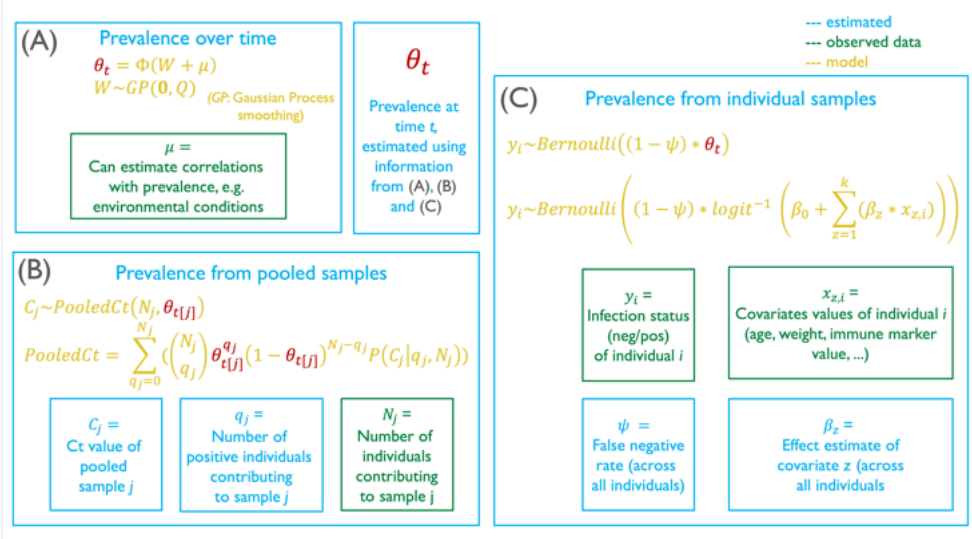

Figure 1: Multilevel model. A model of true prevalence (A) ensures that prevalence is estimated smoothly over time, using information about prevalence from the two other models, which is a shared parameter (highlighted in red). Model (A) is able to test correlations between prevalence fluctuations and other variables such as temperature and precipitation. Model (B) illustrates the model that estimates prevalence from pooled samples, using the Ct value and number of contributing individuals as input data. Model (C) uses individual-level data to estimate prevalence as well as correlates of infection status in a joint likelihood formulation.

Modeling individual samples

Individual (i) test result (negative or positive for biomarker presence) was modeled as a binary variable y_i (0 = negative, 1 = positive), assuming a Bernoulli distribution with a probability of testing positive that is determined by ~~$\hat{\theta}_{t[i]}$~~ $\theta_{t[i]}$ (true prevalence

in the population at time t):

$$y_i \sim Bernoulli\left((1-\psi) * \theta_{t[i]}\right), \tag{1}$$

where ~~$\hat{\psi}$~~ ψ is the false negative rate that accounts for the lower prevalence resulting from the presence of false negative samples. To enable estimation of the correlation between covariates (e.g., biomarkers, age, body weight) and infection status, which can provide ecological insights and may be used to predict shedding status, a regression model is added, resulting in a joint distribution for y_i:

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

where ~~$\hat{\beta}_0$~~ β_0 is the intercept and ~~$\hat{\beta}_z$~~ β_z is the coefficient for covariate x_z out of k covariates. Importantly, this regression model includes ~~$(1-\hat{\psi})$~~$(1-\psi)$, which corrects for false negatives when estimating the covariate coefficients. Without this correction, the presence of false negatives would lead to wrong coefficient estimates. Note that y_i occurs in two parts of the model (equations **??** and **??**) because $\theta_{t[i]}$ is a shared parameter and can therefore not be redefined as the corresponding part of equation **??**, resulting in this joint likelihood formulation.

The probability of an individual being positive, even when testing (false) negative, can be calculated as

$$P_i^+ = logit^{-1}\left(\beta_0 + \sum_{z=1}^{k}(\beta_z * x_{z,i})\right), \tag{3}$$

where individual shedding probability is informed by an individual's covariate value(s). When there is a correlation between shedding status and one or more individual-level covariates, status can be predicted at the individual level, which is helpful for identifying which individuals may have tested false negative.

The prior distribution for ~~$\hat{\psi}$~~ ψ can be a beta distribution as it is bounded by 0 and 1. Because in many cases low false negative rates will be more likely, this could be a slightly informative distribution such as $Beta(1,3)$. The prior distributions for the regression coefficients ~~$\hat{\beta}_z$~~ β_z will depend on the covariate and the way in which their correlation with shedding status is modeled, but in many cases this can be an uninformative normal distribution such as $Normal(0,3)$ when using scaled covariates.

Modeling pooled samples

The goal of this model is to estimate the proportion of positive bats using the Ct value of a pooled sample. The analysis of pooled samples can be challenging, leading to a large body of studies on pooled testing (also called group testing or composite testing, depending on the field) addressing the different problems related to pooled samples [**RN8**, **RN10**, **RN11**]. Most studies have focused on pooled testing in the context of laboratory assay cost reduction, where the main challenge is to find the optimal number of samples to pool given an expected proportion of positives [**RN18**]. An evolving challenge that is more applicable for understanding transmission dynamics is how to estimate the proportion of positive individuals. A number of approaches have been proposed for this, with many based on the model presented by [**RN6**]:

~~$\pi = (1 - (1 - \hat{\theta})^n)$~~ $\pi = (1 - (1 - \theta)^n)$,

where π is the probability that a pooled sample tests positive, ~~$\hat{\theta}$~~ θ is prevalence, and n is the number of samples in the pool. Parameter π can then be used to model $Z \sim Bernoulli(\pi)$, where Z is a binary observed variable indicating whether or not the pooled sample is positive. This implementation has, for example, been proposed as a way to model prevalence dynamics over time for SARS-CoV-2, in combination with individual data [**RN13**]. This approach has two key limitations however. A first is that above a certain combination of pool size and prevalence (around 50%), most pooled samples will be positive, resulting in large uncertainty intervals surrounding the prevalence estimates. A second weakness is that this approach does not account for the fact that the concentration of pathogen is diluted by samples containing a lower concentration, including negative samples. This dilution effect has proven to be particularly difficult to address [**RN9**].

To date, most approaches have used binary test data for estimating prevalence using pooled samples [**RN25**, **RN6**, **RN5**]. Most assays, however, provide quantitative data, which are then turned into a binary negative/positive result based on a threshold value, and the additional information provided by the quantitative assay is lost. This quantitative information offers opportunities, however, that can address both limitations of the binary approach. Although few studies have developed methods to use the full quantitative test results for estimating prevalence from pooled samples [**RN8**, **RN9**], the work by [**RN10**] in particular has shown how promising this approach can be. They used a Bayesian mixture model approach to estimate prevalence, taking into account the dilution effect based on the distribution of biomarker values (e.g. pathogen concentration) of negative and positive samples. ~~These methods seem to work well for standard probability distributions, but currently no solution exists when the underlying distribution of~~

~~values is not standard (e. g. Gaussian). Unfortunately, this is typically the case in biological systems, where biomarker distributions are likely to be skewed and non-standard , and can change over time . Here, we present an approach that can use any distribution of biomarker values to estimate prevalence from pooled samples~~A crucial part of these approaches is the use of a probability density function of positive test values. The methods in [**RN8**, **RN10**] provide a useful approach for estimating these. To complement these approaches, we provide an algorithm to numerically calculate this probability density function so that it covers all possible combinations of numbers of positive and negative individuals while taking into account the underlying distribution of test values in the population.

We modeled pooled samples using their cycle threshold (Ct) value, a measure of the concentration of viral genetic material obtained using qRT-PCR (lower Ct value = higher concentration). The virus concentration in a pooled urine sample is determined by three key factors that influence the final pooled concentration: (1) proportion of positive bats, (2) concentration of virus shed by each positive bat, (3) ~~urine volume~~ relative urine volumes collected from each bat. Here we focus on the first two factors, and assume that the volumes collected from each bat are equal. In order to estimate the proportion of positive bats using the Ct value, it is necessary to calculate a probability distribution of Ct values for pooled samples, as this in turn enables calculating the likelihood of observing certain values given a combination of parameter values. A Ct probability distribution can be calculated by combining two key parts, a standard binomial probability density function (to take into account prevalence) and an ad-hoc distribution of probabilities of observing a pooled Ct value given a combination of negative and positive bats:

$$C_j \sim PooledCt(N_j, \theta_{t[j]}), \tag{4}$$

where C_j is the Ct value of pooled sample j, N_j is the total number of bats contributing to sample j, ~~$\hat{\theta}_{t[j]}$~~ $\theta_{t[j]}$ is prevalence at the time sample j was collected, and

$$PooledCt = \sum_{\hat{q}_j=0}^{q_j=0} {}^{N_j} \binom{N_j}{q_j} \theta_{t[j]}^{\hat{q}_j\, q_j} (1-\theta_{t[j]})^{N_j-\hat{q}_j\, N_j-q_j} P(C_j|q_j, N_j). \tag{5}$$

Here, ~~$\binom{N_j}{\hat{q}_j}\hat{\theta}_j^{\hat{q}_j}(1-\hat{\theta}_j)^{N_j-\hat{q}_j}$~~ $\binom{N_j}{q_j}\theta_j^{q_j}(1-\theta_j)^{N_j-q_j}$ is the binomial probability of observing ~~$\hat{q}_j$~~ q_j positive out of N_j contributing individuals in pooled sample j, given a prevalence ~~$\hat{\theta}_{t[j]}$~~. ~~$P(C_j|\hat{q}_j, N_j)$~~ $\theta_{t[j]}$. $P(C_j|q_j, N_j)$ is the probability of observing Ct

value C_j given ~~$\hat{q}_j$~~ q_j positive out of N_j individuals. ~~$\hat{q}_j$ and $\hat{\theta}_{t[j]}$~~ q_j and $\theta_{t[j]}$ are the estimated parameters, while N_j and C_j are observed. This equation closely matches equation 2 in [**RN9**].

Prior to model fitting, ~~$P(C_j|\hat{q}, N)$~~ $P(C_j|q, N)$ must be calculated for each possible combination of ~~$\hat{q}$~~q, N_j and C_j, which is done according to the following algorithm:

1. Determine all possible combinations (with repetition) of ~~$\hat{q}$~~ q possible Ct values and ~~$N_j - \hat{q}$~~ $N_j - q$ negative values.

2. For each combination:

— 2.1. Transform the Ct values of the positive samples to virus concentrations (conversion based on laboratory controlled testing, or testing of a range of individual samples).

— 2.2. Calculate the mean virus concentration.

— 2.3. Back-transform the mean virus concentration to its corresponding Ct value. Round up the Ct value to the next integer to mimic detection in RT-PCR (a concentration even slightly higher than a certain Ct value will not be detected until the next PCR cycle).

3. Count the number of combinations that result in Ct value C, and divide by the total number of combinations. This is Ct observation probability ~~$P(C_j|\hat{q}, N_j)$~~$P(C_j|q, N_j)$, without accounting for prevalence in the population.

All code used for the calculation of the probability distributions can be found in Supplementary Information.

There are a number of important considerations when calculating ~~$P(C_j|\hat{q}, N_j)$~~$P(C_j|q, N_j)$. A first is that while the algorithm assumes that each Ct value (in step 1) is equally likely, this is rarely the case. The distribution of Ct values in a population rarely follows a uniform distribution, and can instead follow many possible non-standard distributions (e.g., a skewed distribution when low concentrations are more likely). These distributions can also change over time and with changing biological conditions [**RN95**]. When this is the case, probability ~~$P(C_j|\hat{q}, N_j)$~~ $P(C_j|q, N_j)$ can be calculated by first calculating the total probability of each combination, then taking the sum of the total probabilities of all combinations that result in Ct value C, and dividing this by the sum of all total probabilities of all combinations. When the underlying Ct distribution changes over time, or under certain conditions, ~~$P(C_j|\hat{q}, N_j)$~~ $P(C_j|q, N_j)$ must be calculated for each of these situations. Individual samples, if collected, can be used to inform this distribution.

A second consideration is that urine volume is assumed to be equal for all N contributing bats. If this is not the case, the combinations can be corrected by

normalizing for volume in the sample. This step requires knowledge of the volumes contributed by each individual. While this is possible in situations where samples are pooled after collection from individuals, this is unrealistic in field conditions. In this situation, the most parsimonious solution is to assume that all bats contributed equally to a pooled sample. This will of course rarely be the case, but variation in contributed volumes should not affect inference as long as it is not biased. Such biases could arise if infected bats, or bats shedding lower or higher virus concentrations, excrete different volumes than others. It is possible however to account for this when calculating ~~$P(C_j|\hat{q}, N_j)$~~ $P(C_j|q, N_j)$ if there is a model of how this bias occurs.

A third consideration is computational burden, which enforces a limit on the number of Ct values and the number of contributing bats. This is due to the fact that for each possible Ct value of a pooled sample, a probability is calculated for each possible combination $\frac{(C+N-1)!}{N!(C-1)!}$ of Ct values C and individuals N. For example, in a simple situation where only 2 Ct values are possible, and a sample has 3 contributing individuals, the probability of observing a certain Ct value of the pooled sample must be calculated for $\frac{(2+3-1)!}{3!(2-1)!} = 4$ combinations. For more realistic numbers of 15 possible Ct values and 10 individuals, this becomes 1,961,256 combinations, increasing exponentially and rapidly reaching a maximum computationally feasible limit around combinations above 15 Ct values and 15 individuals. There are solutions for this, however. One solution would be to discretize Ct values into larger intervals (e.g., [21-24), [24-26), etc.), and/or setting all numbers of individuals above a certain maximum value equal to that value. This would lower the number of possible combinations and reduce computation time to feasible levels. Another solution, which would not require discretizing biomarker values or limiting the number of contributing individuals, would be to approximate the Ct probability distribution using Monte Carlo simulation/sampling [**RN41**] to generate a large number of random combinations of all values (versus numerically calculating every possible combination). While these solutions are likely to still result in good prevalence estimates, this will depend on the situation and should be tested with simulations prior to model fitting. We recommend taking these pool size requirements into account during the field experimental design process.

A full working example of the procedure to calculate the Ct probability distribution is provided in Figure **??**.

Modeling true prevalence in the population over time

The final model component is a model of ~~$\hat{\theta}_t$~~ θ_t dynamics, which explicitly incorporates the information about prevalence from individual and pooled samples through their respective models. This is possible because prevalence parameter ~~$\hat{\theta}_t$~~ θ_t is shared

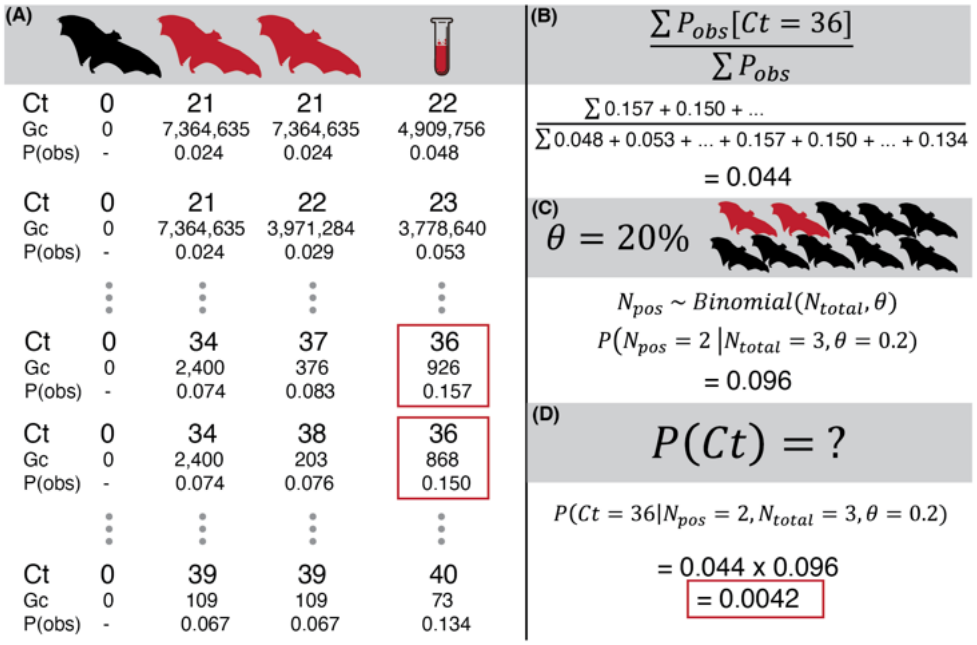

Figure 2: Illustration of how the probability of observing a Ct value in a pooled sample is calculated. In the example, we want to calculate the probability of observing a Ct of 36 with 2 out of 3 positive bats and 20% prevalence in the population. First (A), the pooled Ct value is calculated for every possible combination (with repetition) of 1 negative and 2 positive bats. For each combination, the Ct values (ln scale) are converted to genome copies (Gc) so that the pooled concentration can be calculated on a linear scale. The pooled genome copy concentration is then converted back to a Ct value, rounding up to the next integer to emulate the RT-PCR detection process. Next, for each combination of Ct values the corresponding probability of observing the pooled value is calculated by summing the respective individual probabilities that are estimated from the Ct distribution in individual bats. The probabilities corresponding with the target value of 36 are then summed and divided by the sum of all probabilities, to get an overall probability of observing Ct 36 (B). This probability is multiplied by overall prevalence in the population. The probability of observing 2 out of 3 positive individuals given a prevalence of 20% is then calculated (C) and multiplied by the probability of observing Ct 36 to get the final Ct probability given 20% prevalence and 2 out of 3 positive individuals (D). This example was randomly chosen for illustration purposes, and these steps are repeated for each possible combination of Ct values and contributing individuals.

by both models, which further ensures that each of those two model components can benefit from the information about prevalence contained in the other. Prevalence ~~$\hat{\theta}_t$~~ θ_t changes over time t in a smooth way where prevalence will be more similar for times that are close together than for those that are farther apart. This temporal autocorrelation can be modeled in a variety of ways, and the choice of which model to use will depend on the research questions of interest. If the goal is to estimate prevalence dynamics over time, relatively simple smoothing functions can be used such as splines, weighted average or kernel functions [**RN42**]. If the goal is to model the underlying biological dynamics, it will be necessary to develop a more complex transmission model [**RN43**]. Here, we used a relatively simple Gaussian Process (GP) smoothing function, which uses a Gaussian kernel to model prevalence over time. This approach was based on the one used in [**RN13**].

A GP is a time continuous stochastic process $\{X_t\}_{t\in\tau}$ where the set of variables $X_t = (X_{t1}, ..., X_{tn})^\tau$ is a multivariate Gaussian random variable (i.e., every combination of $(X_{t1}, ..., X_{tn})$ has a univariate Gaussian distribution). Because $\theta_t \in [0, 1]$, a transformation must be used to map the real support of X_t to the $[0, 1]$ interval, for which we used the inverse probit function $\Phi(\cdot)$. We did this by modeling a latent prevalence process $W := \{W_t\}_{t\in\tau}$ and transforming this to prevalence ~~$\hat{\theta}_t = \Phi(W_t)$~~ $\theta_t = \Phi(W_t)$. As prevalence and the form of the unobserved dynamic process are unknown, we used a GP prior on W with a covariance function that enables interpolation of prevalence over time (i.e., smoothing). There are multiple options for suitable covariance functions. Here, we used the exponentiated quadratic covariance function, which includes parameters for both the amplitude (lengthscale ℓ) and the oscillation speed (σ^2) of the smoothing process,

$$Q = Cov(t, t'|\sigma^2, \ell) = \sigma^2 exp\left(\frac{-(t-t')^2}{2\ell^2}\right). \tag{6}$$

W_t thus becomes $W_t \sim GP(\mathbf{0}, Q)$, a zero-mean GP that allows independent modeling of the mean, which is useful for modeling the effect of covariates on prevalence, as ~~$\hat{\theta}_t$ becomes $\hat{\theta}_t = \Phi(W_t + \mu)$~~ θ_t becomes $\theta_t = \Phi(W_t + \mu)$, where μ can be any regression model.

A useful property of the covariance function is that by fitting the lengthscale ~~/amplitude~~ parameter (ℓ), we can learn from the data how ~~temporally correlated prevalence~~ prevalence covaries over time is: the ~~correlation~~ covariance between prevalence values separated by a time interval ℓ will be exactly $\sigma^2 exp(\frac{-\ell^2}{2\ell^2}) = \sigma^2 exp(-\frac{1}{2}) = \sigma^2 0.61$, for an interval of 2ℓ this will be $\sigma^2 exp(-\frac{2^2}{2}) = \sigma^2 0.14$, and so on.

The prior distributions for parameters σ and ℓ can be any continuous positive

distribution. We used a truncated normal distribution for σ ($Normal(0, 1)$, with 0 as lower bound for sampling) and an inverse gamma distribution for ℓ ($Inv - Gamma(2.5, 150)$). All priors used for model fitting can be found in the code in Supplementary Information.

Testing model performance using simulated data

To test how well the model can estimate parameters under various circumstances, we simulated datasets that resemble realistic infection ~~prevalence~~ sampling scenarios. These datasets consisted of individual-level samples (collected directly from captured bats) and pooled urine samples (collected using a sheets under a roost), collected at certain time intervals (e.g., [**RN22**, **RN23**, **RN24**]). We created a main simulated dataset that resembles a common situation with regards to sample size and temporal resolution and was used as a point of reference for all analyses. To test model performance in different scenarios this main dataset was adapted in a number of ways that are described below.

For the main dataset, an autocorrelated fluctuating prevalence time series was generated for a time period of 300 (an arbitrary number, where the unit can be, but is not restricted to, days) time points using a b-spline function with knots at times 1, 100, 200 and 300. Coefficients for the b-spline function were chosen so that the function would result in ~~realistic prevalence fluctuations~~reasonable prevalence fluctuations, not based on a specific system but useful for testing model performance under a range of sample availability scenarios (Figure **??**). Ten sampling sessions were selected to occur evenly between times 1 and 300. At each sampling session, 50 individual-level catch samples and 50 pooled samples were generated. The infection status (negative/positive) of each individual sample was generated using a Bernoulli distribution with success probability equal to population prevalence at the time of the corresponding sampling session (Figure **??**A). To test how well the model was able to account for false negative results in individual-level samples, a proportion (10%) of randomly selected positive bats was changed to negative.

For each pooled sample, a Ct value was generated in four steps (Figure **??**C). First, the number of bats contributing to the sample was simulated using a negative binomial distribution with size 30 and mean 2.3 (which results in a range between 1 and 10, with most numbers around 1 to 4). Next, each of the contributing bats was randomly assigned an infection status using a binomial distribution with success probability equal to prevalence at the corresponding sampling session. Then a Ct value was generated for each individual bat, with negative bats receiving a Ct value of 0 and positive bats receiving a Ct value randomly drawn from a non-standard, realistic probability distribution of Ct values. Last, the resulting Ct value of the pooled sample was calculated by first converting each individual Ct value to number

of genome copies [**RN95**], calculating the mean number of genome copies (including the negative samples), and re-converting this mean of the pooled sample to a Ct value. 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 [**RN9**, **RN10**].

To test whether the model is able to estimate the effects of covariates on the infection status of an individual, we simulated four covariates with information at both the individual and the population level. We simulated individual covariates with a range of association strengths with infection status (Figure **??**B). This was done by randomly drawing a value for each covariate from a normal distribution for each individual, where the mean depended on infection status. For the strongly correlated covariate the means were 1.5 units apart for negative and positive individuals, with standard deviations of 0.5 (resulting in a regression coefficient of 5.3 log odds). For the moderately correlated covariate the means were also 1.5 units apart but the standard deviation was 1 (resulting in a regression coefficient of 1.9 log odds). For the covariate that did not correlate, the mean was 0 and the standard deviation was 1 (resulting in a regression coefficient of -0.11 log odds). Main dataset simulation parameters are summarized in Table **??**. Additionally, we show the importance of accounting for false negative individual samples when estimating covariate effects by fitting a model that does not include the false negative rate parameter.

Last, to test model performance under different scenarios of data availability, we generated additional scenarios that are outlined in Table **??**, including ~~biological~~ examples of when these scenarios can occur. Details and results for these scenarios are provided in Supplementary Information, including combined scenarios.

Model implementation and code

All coding was done in R [**RN16**]. Model fitting was done with Stan [**RN14**] using R package rstan [**RN15**]. Plotting was done using packages ggplot2 [**RN98**], ggridges [**RN105**], patchwork [**RN99**] and Rcolorbrewer [**RN100**]. Prevalence splines were generated using the package splines [**RN16**]. Ct value probability distribution generation used the package Rccpalgos [**RN101**]. Supplementary information (including all code) is available online at `https://doi.org/10.5281/zenodo.10660032`.

Table 2: Overview of the parameters used for the main simulated dataset

Parameter	Value
Times	300 (arbitrary) time units
Number of sampling sessions	10
Timing of sampling session	Every 34 time units
Individual samples per session	50
Pooled samples per session	50
Infection data of individual samples	Binary (negative or positive)
Infection data of pooled samples	Concentration (Ct value)
False negative rate	10% of positive individual samples tests negative
Individual covariate, strong correlation	Effect estimate = 5.3
Individual covariate, weak correlation	Effect estimate = 1.9
Individual covariate, no correlation	Effect estimate = -0.11
Ct distribution used to simulate pooled Ct values	Skewed low to high (details in Supplementary Information)

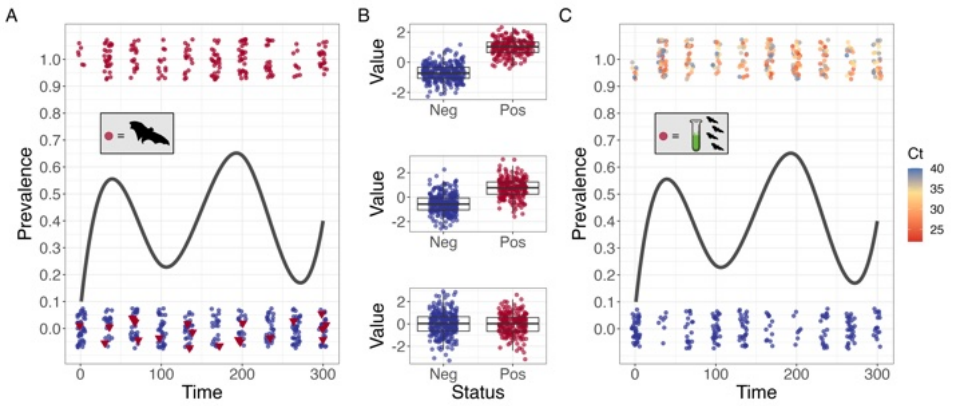

Figure 3: Simulated data. Black lines show true prevalence in the population, which was used to generate samples for 10 sessions over a period of 300 time points. Panel A shows individual negative (blue) and positive (red) samples, with false negative samples shown as red triangles. Panel B shows boxplots and data points for three simulated covariates for individual samples, with correlations being strong (top), moderate (middle) and random (bottom). Panel C shows pooled negative (blue) and positive (blue to red gradient corresponding with Ct value) samples. Note that infection data are binary (neg/pos) for individuals, and concentrations (Ct values) for pooled samples.

Table 3: Simulated scenarios to test model performance. Details and model fit results are provided in Supplementary Information.

Scenario	Details and examples
Pooled data only	Only pooled samples are used for model fitting. $\hat{\psi}$ ψ not estimated. This situation occurs when it is not possible not collect any individual data. Example: collecting wastewater samples for COVID-19 monitoring [**RN19**].
Individual data only	Only individual samples are used for model fitting. $\hat{\psi}$ ψ not estimated. This situation is the most common when sampling populations. Example: cross-sectional sampling for monitoring arenavirus prevalence in rodents [**RN85**].
Irregular sampling	The timing of sampling sessions is not regular, resulting in uneven time gaps between sessions. This situation occurs when regular sampling is not possible, or when sampling sessions need to be canceled due to conditions. Example: gaps in influenza A monitoring time series due to political instability [**RN87**].
Low sample sizes	Lower sample sizes (20 instead of 50 for each sample type) per session. This situation occurs when it is not possible to sample many individuals. Example: logistically challenging captures of lions for canine distemper virus monitoring [**RN89**].
Ct distribution mismatch	The distribution of Ct values used to calculate the likelihood for pooled sample Ct values is different from the true distribution used to simulate pooled Ct values. This situation can occur when the distribution of Ct values in the population is not well known. Example: small numbers of positive samples in individual bats make it difficult to describe the viral load distribution of filoviruses [**RN90**].
Pool contribution count error	An error is added to the number of individuals contributing to a pooled sample. This situation occurs when it is difficult to count or estimate the number of individuals contributing to a pooled sample. Example: environmental sampling for *Leptospira* sp. prevalence estimation [**RN91**].
Prevalence dynamics shape	A number of different, uncommon prevalence fluctuations are used to simulate the data. This situation occurs because prevalence dynamics can vary strongly depending on many factors. Example: measles prevalence dynamics exhibiting multi-annual cycles of varying magnitude [**RN92**].
Prevalence covariate	A covariate that correlates with prevalence is estimated using Gaussian Process regression. Example: climate can drive inter-annual cycles of cholera transmission [**RN97**].

Results

Shedding prevalence dynamics estimated using the combined pooled and individual data, closely matched the true dynamics, with true prevalence always within the ~~95% credible intervals~~ posterior distribution (Figure **??**A). All individual and prevalence covariate coefficients were estimated correctly (Figure **??**B-D). The ability to predict individual shedding status depended on the amount of information provided by the covariate: prediction accuracy was 94.2% (95% CrI: 92.8-94.2%) when using the strongly correlated covariate, 81.0% (95% CrI: 79.8-81.8%) for the moderately correlated covariate and 58.4% (95% CrI: 55.6-58.4%) for the random covariate (Supplementary Information). The model correctly estimated the false negative rate ($\hat{\psi}$~~ψ~~) regardless of which covariate was used (Figure **??**C). Moreover, the model was able to identify 100% (21/21) (95% CrI: 100-100%) of individual false negative samples when using the strongly correlated covariate, 76.2% (95% CrI: 66.7-81.0%) when using the moderately correlated covariate and 0.0% (95% CrI: 0.0-9.5%) when using the random covariate. Exclusion of the false negative rate parameter from the model resulted in wrong estimation of the strongly correlating covariate, where the correct coefficient was 5.1 but the posterior mean estimate was 2.8 (95% CrI: 2.4-3.2).

When using only a single type of data (either pooled or individual data) the model was still able to capture true dynamics (Figure **??**A-B), although there was a slight overestimation of prevalence when only using pooled data, and a slight underestimation when using individual data only. The false negative rate could not be estimated in the absence of pooled data as there was no additional source of information to provide information about true prevalence over time.

When sampling sessions were timed irregularly, prevalence dynamics were still estimated well but with a higher degree of uncertainty was observed between larger time gaps (Figure **??**C and Supplementary Information) and for the times beyond the time limits of the data. For regular sampling with low sample sizes the model still performed well (Figure **??**D). Asynchronous sampling of pooled and individual sessions resulted in prevalence dynamics that were very similar to those of the "main" simulated dataset (Supplementary Information). When combining irregular and asynchronous sampling with lower sample sizes or with fewer sampling sessions prevalence dynamics were estimated well, but with higher degrees of uncertainty due to the lower sample size or during larger time gaps without available samples (Supplementary Information).

A mismatch of the Ct distribution in the population (i.e., the Ct distribution used to construct ~~$P(C_j|\hat{q}, N_j)$~~ $P(C_j|q, N_j)$ did not correspond with the distribution used to simulate Ct values for pooled samples, see Supplementary Information for

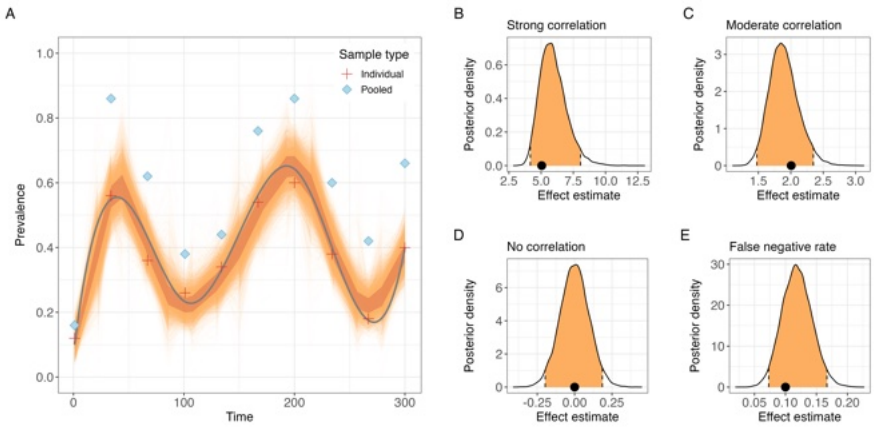

Figure 4: Model outputs for the main simulated dataset. (A) shows the distribution of fitted prevalence curves (cloud of 6,000 iterations from 5 chains) with 50% credible interval band overlaid. The ~~black~~ grey line is the true prevalence. The proportion of positive pooled and individual samples in each sampling session is shown using diamond and plus shapes, respectively. Panels (B) to (D) show the posterior distributions (95% credible intervals in orange) for the three different covariates, and panel (E) shows the posterior distribution for the false negative rate ~~ψ~~ $\hat{\psi}$, with black dots indicating the true values.

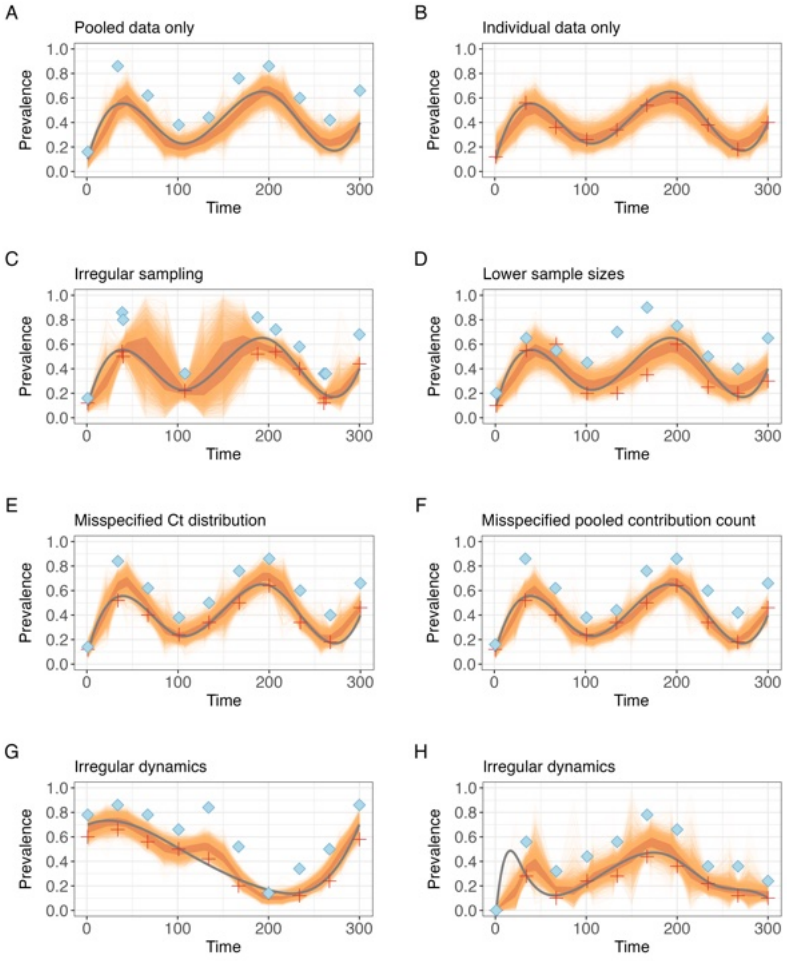

Figure 5: Fitted prevalence curves for different simulated scenarios. All scenarios use the same sample types, sizes and sessions as the main scenario shown in Figure 2, except where indicated. Session prevalence of pooled and individual data is shown using diamond and plus shapes, respectively. Black lines show true prevalence in the population. Specifics for each scenario are: (A) only pooled data; (B) only individual data; (C) sampling sessions are unevenly spaced over time; (D) lower sample sizes (10 samples per sample type instead of 30) per session; (E) the Ct distribution used to simulate Ct values of pooled samples was not the same as that used to calculate the Ct probability distribution in the model, with the shapes inverted (i.e. low Ct values more likely); (F) an incorrect number of individuals contributing to a pooled sample was provided to the model for 50% of pooled samples; (G) and (H) data were simulated using irregular, unconventional prevalence dynamics.

details) had a noticeable effect on the estimated prevalence dynamics (Figure **??**E). Specifically, the model tended to overestimate prevalence, particularly during peaks, despite overall good performance. This effect was less pronounced when the distribution was less different from the true distribution (Supplementary Information). The model was not sensitive to moderately misspecified counts of the number of individuals contributing to a pooled sample (Figure **??**F; 30% of the data were off by N = 1, 20% by N = 2), but was more strongly affected by large misspecifications (80% wrong by 1 or 2, 80% wrong by 1 to 5; Supplementary Information).

Last, the shape of the prevalence dynamics did not affect the model's ability to estimate prevalence, as long as data were available to inform the fluctuations (**??**G and Supplementary Information). For example, the dynamics in Figure **??**H show an initial peak that was not predicted by the model because this peak occurred between two sampling sessions.

Discussion

Sample pooling offers major benefits through collecting data from multiple individuals at the same time, lowering costs for collection and testing, and enabling the use of samples that would otherwise be disregarded (such as sewage or fecal/urine under bat roosts or in animal dens) [**RN19**, **RN21**, **RN23**]. This study presents a Bayesian modeling approach that enables the estimation of prevalence dynamics from both pooled and individual samples by leveraging infection concentration of infectious agent in the pooled samples, allowing the distribution of infection ~~concentration~~ concentrations to be any shape, and accounting for false negative results.

The model is able to successfully reconstruct prevalence dynamics for a wide range of eco-epidemiological scenarios. Model performance was tested for a range of ~~realistic~~ relevant scenarios of infection dynamics and sampling schemes including irregular prevalence fluctuations, irregular timing of sampling, inclusion of associated individual-level covariates, ~~and~~ misspecified counts of individuals contributing to the pooled samples, and combinations of multiple scenarios. The model performs well when only one sample type was provided, which is particularly encouraging in the case of pooled samples, as it shows that field studies targeting only pooled samples would still allow precise reconstruction of prevalence dynamics. These results highlight the key strengths of the model: the explicit modeling of the mixing process in pooled samples allows accurate estimation of prevalence even when using only pooled samples, and the inclusion of pooled samples also enables correcting for the prevalence estimation bias in individual samples introduced by false negatives when both sample types are available. False negative rate is an epidemiological parame-

ter commonly neglected in wildlife studies, yet important for inferring dynamics of infection at the individual level. In the model, estimation of false negative rates is made possible by the explicit integration of information about prevalence included in both data types. Importantly, accounting for false negative results ensures that covariate coefficients in individual-level regression models are estimated correctly, which we show would otherwise lead to estimation errors.

The ~~main technical advance of the model is the calculation of a~~ model introduces an algorithm to empirically calculate the probability distribution of observing a certain infection biomarker (here Ct from qRT-PCR) value given the estimated prevalence in the population and the number of individuals that contributed to the sample. This probability distribution enables the calculation of a likelihood for the biomarker values of the pooled samples. This approach for calculating a probability distribution can be adapted to other systems (e.g., analyzing pooled SARS-CoV-2 samples for monitoring prevalence) and other biomarkers (e.g., antibodies, blood chemistry). ~~It builds on existing methods for the analysis of pooled samples and advances them by providing a solution for dealing with the dilution effect where biomarker concentration in the pooled sample decreases or turns negative due to mixing with negative or lower-concentration samples . The approach also provides a way to deal with~~ The approach can incorporate any non-standard family ~~distributions~~ distribution of the biomarker~~, thereby making it more flexible than existing methods~~ . Encouragingly, we found that the model is quite robust against misspecifications of the underlying biomarker distribution. The calculation of this probability distribution function relies on a correct determination of the distribution of biomarker values in the population. We found that assuming a distribution that differs strongly from the real distribution can result in biased prevalence estimates. We therefore recommend an in depth prior exploration of biomarker distribution in the population, as well as a sensitivity analysis to assess how different realistic shapes of the distribution affect model output.

Prevalence reconstruction is a goal for many epidemiology and disease ecology studies, but this is often done as a necessary step towards learning what the drivers of pathogen transmission are. Such drivers can be intrinsic, such as individual immunity, herd immunity, individual variation in shedding, or behavior/movement (which can affect contact/transmission rates), or extrinsic, such as temperature and rainfall affecting pathogen survival, food availability affecting individual stress (which in turn affects immune competence, susceptibility and shedding). The modeling framework provides a way to incorporate and statistically test the effect of such covariates on the individual and the population/prevalence level. This enables testing of hypotheses about intrinsic or extrinsic drivers of infection, thereby contributing to a more mechanistic understanding of infection dynamics, beyond the phenomenological patterns. This also enables the development of models to predict

prevalence.

The current model formulation has a number of requirements. Firstly, the model uses estimates of the number of individuals that contributed to a pooled sample. While the model is robust against moderately misspecified counts, we find that errors have to be within reasonable limits. However, when these counts are unknown or uncertain, this can be incorporated in the model by specifying a prior distribution of the number of individuals contributing to a pooled sample based on available data. A second model requirement is that the distribution of biomarker values, which are used to calculate the biomarker probability distribution of pooled samples, is assumed to be constant over time. Although this can be a reasonable baseline assumption, recent work suggests this may not always be the case [**RN95**]. Therefore, it is possible to adapt the model using a time-dependent probability distribution when pathogen shedding concentrations are known or suspected to be higher during certain periods. We recommend an in-depth analysis of the distribution of biomarker values in ~~individuals~~ wild individual samples over time to determine whether the probability distribution used in the model needs to be time-dependent.

The model presented here provides a way to simultaneously leverage pooled and individual samples to accurately estimate the true underlying prevalence of infection in a population. It introduces a way to explicitly account for the biological mixing/dilution process in pooled samples, and ensures that individual covariate effects can be estimated correctly when false negative results are possible (this requires the use of both pooled and individual samples). The model is also shown to be robust against common issues associated with field-based data collection, such as observation noise and the often unknown shape of the underlying prevalence fluctuations. Crucially, this approach enables the accurate reconstruction of prevalence dynamics even when using pooled samples only, which is encouraging for designing lower-cost sampling strategies. The application of this model can directly enhance the efficacy and efficiency of bio-surveillance efforts by increasing inference and prediction. This is of particular interest in the case of wildlife that hosts pathogens of concern for human and animal health in geographical areas of high spillover risk.

Acknowledgments

This work was funded by the U.S. Defense Advanced Research Projects Agency (DARPA) PREEMPT program Cooperative Agreement (D18AC00031), and the U.S. National Science Foundation (DEB-1716698; EF-2133763). The content of the information does not necessarily reflect the position or the policy of the U.S. government, and no official endorsement should be inferred.