Bayesian data integration for ~~small area~~ estimation of ~~pathogen~~ `coronavirus` prevalence dynamics from pooled ~~and individual~~ data

June 28, 2022

Abstract

Estimating the prevalence of a disease, `, such as coronavirus,` is necessary for evaluating and mitigating risks of its transmission ~~within or between populations~~. Estimates that consider how prevalence changes with time provide more information about these risks but are difficult to obtain due to the necessary survey intensity and commensurate testing costs. `Motivated by datasets on SARS-CoV-2 at Notre Dame and coronavirus surveillance in African bats,` we propose pooling and jointly testing multiple samples to reduce testing costs and use a novel nonparametric, hierarchical Bayesian data integration model to infer population prevalence from the pooled test results. This approach is shown to reduce uncertainty compared to individual testing at the same budget and to produce similar estimates compared to individual testing at a much higher budget through three ~~synthetic~~ `simulation` studies and two case studies of natural infection data. Furthermore, the data integration procedure enables estimates of prevalence dynamics for sub-populations of the study ~~similar to small area estimation~~ by combining information from pooled and individual data.

Keywords: small area estimation, survey methods, Bayesian methods, time series, Gaussian process

Statement of Significance

Combining pooled samples with individual results can give more precise estimates of disease prevalence over time. In particular, our ~~novel~~ approach leverages information from pooled samples and individuals from other subpopulations to give better estimates. While the work is motivated by ~~disease~~ `coronavirus` dynamics, the methods could be applied to any binary process over time.

1 Introduction

`During the early phases of the COVID-19 pandemic many public health policies were tied to the prevalence of infectious individuals in the population. Similarly,` the risk of pathogen transmission from animals to people (zoonotic spillover) is a function of the prevalence of the pathogen in animal populations in ~~space and~~ time [1, 2]. Therefore, understanding patterns of pathogen prevalence is `a` fundamental `problem` ~~to mitigating zoonotic disease events~~ but the intensive testing required to do so is often prohibitively costly. Testing and survey costs can be reduced by testing pools of samples, but it is difficult to estimate prevalence from pooled samples over time. In this paper, we develop inferential tools for efficiently estimating pathogen prevalence `over time` ~~time series processes~~ by combining pooled and individual data, which also enables ~~small area~~ estimation of prevalence within subpopulations. These methods `are motivated and` ~~can be~~ applied to estimate `coronavirus prevalence in human and bat populations.` ~~and mitigate the risk of zoonotic spillover and therefore provide a tool for pandemic prevention.~~

~~Prevalence estimation is central to disease surveillance in reservoir host populations [3], however, because true prevalence is not observed except in trivial settings, estimating prevalence requires sampling, testing, and inference. Modeling prevalence dynamics can lead to more precise predictions of prevalence and, thereby, spillover but most approaches require intensive sampling and testing efforts [4]. Furthermore, the diversity~~

and complexity of additional factors that contribute to spillover risk limit the availability of resources dedicated to sampling and testing for prevalence estimation [1]. Efficient allocation of testing resources coupled with dynamic modeling will enable better identification of current and future pathogenic hazards. A popular strategy for reducing testing costs when prevalence is expected to be low is to test multiple samples jointly through a procedure known as group testing or pooled testing [5]. Pooled testing proceeds by randomly constructing groups or pools of individual samples and expending a single test on each pool. Testing costs are generally proportional to the number of laboratory tests conducted rather than the number of individuals tested; hence, pooled testing can reduce overall costs relative to testing specimens individually. With a perfect test (sensitivity and specificity = 1), a positive result implies one or more individuals in the pool is infected, and a negative result implies all individuals are not infected. These implications are complicated slightly by imperfect tests and pooling-induced dilution, though these limitations are usually outweighed by benefits in practice. Individual diagnoses could then be obtained by retesting members of pools that test positive. Originally developed to scale up the US military's ability to screen recruits for syphilis, the strategy has seen a resurgence in popularity due to the COVID-19 pandemic [6]. When applied to syphilis screening and standard COVID-19 testing, group testing is used in a diagnostic capacity. However, when individual results are not required and pathogen prevalence is the quantity of interest, as is the case in surveillance of reservoir host populations, follow-up testing can be omitted, further reducing testing costs. [7] found that pooled testing can be used to estimate prevalence as accurately and more efficiently than individual testing. However, their results are not directly applicable to observations that exhibit natural temporal correlation, such as multiple samples collected from the same site at different times.

Knowledge of the distribution of infection with respect to time enables better understanding of the factors that lead to high prevalence. However, survey limitations often beget inferences with limited temporal scope or resolution and wide uncertainty intervals. Efficient sampling and inference strategies for estimating prevalence therefore must be extended to account for and leverage the correlation among temporally-indexed observations. Such an extension should achieve interpolation between observation times and smoothing of noisy observations. Interpolation generalizes pointwise inferences to unobserved, intervening times and enables improved qualitative and quantitative characterizations of trends that may not be apparent from pointwise inferences alone. Independent estimates of prevalence at different points in time are noisy; under the assumption that nearby events (in time) are more closely related than distant ones, smoothing de-noises estimates by sharing information between neighboring observations. Smoothing can also produce more comprehensive representations of uncertainty. In the absence of direct observation, prevalence is modeled as a latent process with a transformed Gaussian process (GP) prior and is related to pooled test results through a Bayesian hierarchical model. GPs have been used to model clinical time series data [8] and to estimate parameters in compartmental infectious disease models [9], but we are not aware of their use in estimating prevalence dynamics in pooled testing scenarios. Autoregressive models are common for time series modeling, going all the way back to the 1960s [10], but we opt for a GP due to the ability to treat time in a continuous manner. Many of our applications, such as capturing and sampling bats do not necessarily occur in discrete time, but this modeling choice does result in additional computational burden.

Small area estimation is a common problem in survey statistics [11]. Traditionally, small area estimation has focused on "how to produce reliable estimates of characteristics of interest such as means, counts, quantiles, etc., for areas or domains for which only small samples or no samples are available, and how to assess their precision" [12]. Small area estimation has also been applied to problems focusing on estimation of time series trends [13]. In the context of pathogen dynamics, it is necessary to understand differences in disease prevalence across subgroups of the population. Often these subgroups have a small number of samples or there is sparse temporal coverage, which can be viewed through the lens of small area estimation. While small area estimation often refers to small geographic areas, in this work we are concerned with subgroups of the population. Bayesian procedures for small area estimation often employ a hierarchical structure for sharing information [14, 15]. Similarly, we will use a Bayesian hierarchical model to share information between pooled and individual samples.

Our goals are threefold: 1) establish the ability of pooled data to recover true, underlying time-varying prevalence, 2) demonstrate that pooled testing can reduce testing costs without degrading estimates or,

alternatively, generate more precise inferences at a fixed testing budget, and 3) develop a data integration procedure to identify pathogen dynamics for subpopulations using a combination of pooled and individual data. The remainder of the article is organized as follows. In Section 2, we introduce the motivating data and the hypothetical pooling design. In Section 3, we briefly review pooled testing and GP methodology, introduce the modeling framework, and comment on relevant computational considerations. Section 4 reports results from simulation studies demonstrating recovery of true prevalence and case studies evidencing the relative efficiency of pooled data and integrating pooled data with individual data. The results from the motivating case studies are reported in Section 5, along with additional consideration of sampling variability inherent to our studies. In Section 6 we discuss implications of the present work and future directions.

2   Data

We consider two datasets that motivate our case studies. The first is an example of pathogen surveillance among a reservoir host population where low-cost prevalence estimation is necessary for determining trends and designing interventions. The second considers disease surveillance in a human population where knowledge of prevalence may instead be used to inform institutional and public health policies.

2.1   Case study: Congo Basin bats

[16] report the results from 3,561 coronavirus surveillance tests performed on samples from bats, rodents and primates over 12 years collected throughout the Republic of Congo and Democratic Republic of Congo. We restrict our attention to samples collected from bats. Sampling efforts occurred roughly monthly, and each site was visited roughly twice per year. An approximately one and a half year lapse in sampling occurred between 2014 and 2015. The general pooling framework is evaluated on a subset of the total sampling interval comprising 752 individual tests performed between August 24, 2015 and July 20, 2018. If an individual was tested twice (e.g., using both fecal and saliva samples), we record it only once and deem an individual positive if any of the multiple samples tests positive. Though the date that each sample was collected is recorded, sampling efforts span multiple days. Because we do not expect prevalence to change meaningfully over the short window of days within which sampling occurs, we aggregate results from the same location that occur within < 10 days of each other and record the date as the date of the first sample collection date in the sampling effort. This procedure is consistent with the description of sampling provided by the authors and reflects the reality of sampling reservoir hosts. All processing described here was performed in advance of any analysis and is identical between pooled and individual testing analyses.

A second analysis using the data integration procedures is applied to the portion of the data proceeding the 1.5 year lapse in sampling. This portion of the data spans from September 22, 2011 to March 27, 2014. There are a few extended stretches without any survey effort. Our modeling framework contains high levels of uncertainty over this period, as should be expected. This portion of the data contained a total of 593 individual tests. We identified 5 species of interest with the highest temporal frequency of sampling. The remainder of the tests were grouped into pools of approximately size 3.

2.2   Case study: asymptomatic testing at Notre Dame

In the face of the COVID-19 pandemic, many higher education institutions in the United States implemented on-campus diagnostic and surveillance testing during the 2020/2021 academic year. This, in combination with other public health precautions, was intended to limit transmission of SARS-CoV-2 through early detection and subsequent isolation [17]. The University of Notre Dame, a medium-sized university in Indiana, was one such institution. We examine results from nasal swab and saliva tests administered to asymptomatic students on campus on a daily basis between August 3 and December 16, 2020. Data are published publicly in dashboard format on the university's website [18]. The 81,872 tests identified 248 positive cases, for an overall positive rate of 0.30%. In this scenario, individual diagnostic results were sought after, as is often the case when testing human populations. However, prevalence estimation remains relevant. In humans, knowledge of population prevalence is used to inform decision-making (e.g., whether to revert to online instruction). Pooled test results from initial testing may also be used to direct later, follow-up testing to obtain individual

diagnoses. For example, testing bottlenecks may be combated by prioritizing pooled testing and prevalence estimation for decision-making, and follow-up testing of positive pools can occur when testing resources become available.

2.3 Hypothetical pooling design

Both data sets contain individual test results. To evaluate methods for estimating true prevalence from tests performed on pooled samples, we construct hypothetical pools by randomly grouping individual results and computing pooled test results. A pool is positive if one or more individuals tested positive in the original data and negative otherwise, thereby censoring individual-level data.

Broadly, testing costs can be broken down into a fixed cost that is invariant to the number of tests and a variable cost that scales with the number of tests performed. Reductions in variable testing costs can be achieved through two parallel approaches: 1) reducing the number of tests needed to obtain similarly precise and accurate estimates (obtaining the same results at lower cost), or 2) improving the precision and accuracy of estimates at a fixed budget (obtaining better results at the same cost). The first approach is evaluated by comparing prevalence estimates obtained from pooling to prevalence estimates obtained from the original data. The second approach is evaluated by comparing pooled estimates to estimates obtained from a subsample of the original data representing the same number of tests (i.e., fixed variable cost). Let m be the number of individuals in a given pool. Prevalence estimates that use all available data are labeled $m = 1$ and are regarded as the best approximation to true prevalence and therefore the best known alternative to the proposed method. Estimates from subsampled, individual tests are labeled $m = 1^*$, and estimates from pool tests are labeled with their respective pool sizes (e.g., $m = 5$). Synthetic pooling and subsampling is performed within sampling instances or time steps. Pools are constructed using random sampling without replacement. For example, if 17 individuals were originally sampled at time t_i, corresponding to the ith time point in the study, then the $m = 1$ estimate is based on all 17 individuals, the $m = 3$ estimate is based on 5 pools of size 3 and 1 pool of size 2, and the $m = 1^*$ estimate is based on 6 subsampled individuals, because 6 pools and therefore 6 tests are used in the pooling setup. When multiple pool sizes are considered simultaneously, the larger testing budget is used to construct the $m = 1^*$ subsample. In the Congo Basin bat study, pool sizes 3 and 5 are used. In the Notre Dame study, pool sizes 5 and 10 are used.

For the Congo Basin bat study we also demonstrate the ability of our small area estimation procedure to integrate pooled information with individuals to estimate overall prevalence as well as prevalence within different species. Similar to the previous description, pools are synthetically created; however, this type of data pooling procedure is common in disease ecology field studies. Data collection procedures often focus on two types of data collection. The easiest type of data to collect is to collect urine or feces from a collection of individuals. This data is naturally pooled without information on the subpopulations. The pooled data can be augmented with catching individual animals. This catching process is considerably more costly, but can provide information relevant to identifying subpopulations and relationships with associated covariates.

3 Model Formulation

The methods described here can be used to infer the population prevalence or probability that an individual selected randomly from a specified population is infected with an infectious disease of interest, which is denoted p, at any given time within a specified interval. We observe only whether or not a sample tests positive for the pathogen and the sample collection date. Under appropriate conditions, testing pools or groups comprising samples from multiple individuals confers considerable reductions in testing costs. Therefore, a hierarchical model is used to relate the binary or count data obtained through pooled testing to the probability that a pool tests positive, π, and to relate this pool probability back to individual probability, p. Furthermore, by combining pooled information with individuals we can use the pooled information to inform an overall prevalence which enables more precise estimates for subpopulations.

3.1 Population prevalence and pool probability

In possession of test results from pooled samples, estimating prevalence at the pool level is straightforward—one could, for example, employ a beta prior and binomial likelihood and proceed with a conjugate analysis

analogous to the procedure described in Section 2.2.1 of [7]. However, understanding prevalence of a pool is not typically the goal but rather when prevalence at the individual level is the quantity of interest, any inference on π must be related to p. To do so, we use the inverse prevalence transformation of [19] :

$$\pi = 1 - (1-p)^m, \tag{1}$$

where m is, again, the pool size. Inspection reveals that pool probability is high when either prevalence is high or pool size is large. At lower prevalence, p is approximately linear in π. For high prevalence, however, p is highly non-linear in π at which point many values of p imply similar, large values of π, hindering identifiability. The non-linearity also grows with m (see [19], Figure 1). High prevalence requires small pools; low prevalence permits large pools. [20] describes survey considerations for optimal pool size; [21] additionally considers optimality under two-stage designs, where individuals from positive pools are retested.

3.2 Latent Gaussian process

GPs are popular tools for nonparametric regression as they make no assumptions about the functional forms of relationships and few assumptions generally. They can therefore be used as priors over functions to model unknown, possibly non-linear functions automatically. ~~By definition, a GP is a stochastic process $\{X_t\}_{t \in \mathcal{T}}$ where any select variables $X_t = (X_{t_1}, \ldots, X_{t_n})^\intercal$ are distributed as multivariate Gaussian. The support of a GP is therefore in the real numbers, limiting applications. However, transformations can be used to map this support to different intervals, greatly broadening its uses.~~ The Gaussian distribution function (inverse probit) $\phi(\cdot)$ is ~~one such transformation,~~ commonly used to map GP random variables to $[0, 1]$ for use in classification or to estimate binary probabilities. We adopt a similar approach`, where`~~Because~~ $p_{t_i} \in [0, 1]$~~, it is convenient to model~~a latent prevalence process $W := \{W_t\}_{t \in \mathcal{T}}$ with real support `can be transformed`~~and transform~~ to obtain $p_t = \phi(w_t)$. ~~Prevalence, and therefore the form of the latent process, are unknown.~~ Together with assumptions of smoothness, continuity, and real support, this solicits a GP prior on W. The only additional user-specified components of a GP are the covariance function and priors over covariance hyperparameters. The form of the covariance function determines the nature of the relationship between process values as a function of time, and proximity in time determines the strength of the covariance between any two process values. ~~In general, we expect process values at similar times to be more similar than process values at distant times. This agrees with our notion of smoothing. Covariance can be computed between any locations to obtain a function value, achieving interpolation.~~ A common, general-purpose covariance function, and the function we use throughout this paper, is given by

$$Cov(t, t' \mid \theta) = \sigma^2 \exp\left(\frac{-(t-t')^2}{2\ell^2}\right), \tag{2}$$

known as the exponentiated quadratic covariance function. The hyperparameters $\theta = \{\sigma^2, \ell\}$ control the ~~oscillation speed and amplitude~~ `variance and effective range` of a sampled process, respectively. ~~Many other covariance functions can be used instead to match the application.~~ The covariance matrix C of our multivariate Gaussian distribution is obtained by letting $C_{ij} = C_{ji} = Cov(t_i, t_j \mid \theta)$, and we write $W_t \sim \mathcal{GP}(\mathbf{0}, C_\theta)$.

Here, we have specified a zero-mean GP; a useful property is that if $F \sim \mathcal{GP}(\boldsymbol{\mu}, C_\theta)$ and $G \sim \mathcal{GP}(\mathbf{0}, C_\theta)$, then $F = G + \boldsymbol{\mu}$. The mean can be modeled independently either as a scalar or some function of covariates and the covariance matrix will capture any residual structure not modeled in the mean. Our ~~small area~~ estimation procedure will permit a hierarchical treatment of μ across the subpopulations.

3.3 Sampling models

Models of test results from pooled samples depend primarily on the size of the pools, because the probability that a pooled sample tests positive is a function of both population prevalence and pool size. Different models arise depending on how pool sizes are assigned. At one extreme, pool sizes and number of pools are determined in advance of sampling individuals and all pools are the same size. At the other extreme, pools of many different sizes are tested, possibly in advance of any analysis. Here, we consider three scenarios representing 1) the general case, where all pools may be different sizes, 2) the ideal case in which all pools

are the same size, and 3) an efficient alternative to the general case that represents many realistic situations. Idealized and efficient general models can be regarded as special cases of the general model. In all scenarios, the pool sizes m, number of pools k, and number of individual samples n are assumed to be known for all pools. Unknown pool sizes could, in principle, be estimated from data, but strong prior information on pool sizes or other forms of regularization would be required for identification. In all scenarios, small notational simplifications are obtained when n (and therefore k and m, also) is constant in time.

3.3.1   Individual Estimation

For individual samples and no pooling, the model used to infer prevalence is specified as

$$Z_{t_i}|p_{t_i} \overset{\text{iid}}{\sim} \texttt{Binomial}(k_{t_i}, p_{t_i}), \tag{3}$$

where k_{t_i} is the number of individuals tested at t_i and $Z_{t_i} \in \{1, \ldots, k_{t_i}\}$ is the observed number of infected individuals. Covariates can be included in this model framework, or those that follow, by using a link function where $g(p_{t_i}) = x_{t_i}\beta$.

3.3.2   General Pooling

In the general setting, each pool is given its own probability—no assumptions are needed about the size or number of pools. We use $j \in \{1, \ldots, k_{t_i}\}$ to index pools at time t_i. The sampling model is

$$
\begin{aligned}
Y_{t_i,j}|\pi_{t_i,j} &\overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_{t_i,j}) \\
\pi_{t_i,j} &= 1 - (1 - p_{t_i})^{m_{t_i,j}},
\end{aligned} \tag{4}
$$

where $Y_{t_i,\cdot}$ is a vector of binary variables indicating which pools of samples test positive at time t_i. At each time, $\pi_{t_i,j}$ is indexed by pools, but p_{t_i} is constant. We use this model in the Congo Basin bat surveillance example. If $m_{t_i,a} = m_{t_i,b}$ for all $a, b \in \{1, \ldots, k_{t_i}\}$, this model simplifies to the idealized model described next.

3.3.3   Ideal pooling

In the ideal setting, all pools at time t_i contain m_{t_i} individual samples, and k_{t_i} pools are tested. This requires $m_{t_i} \times k_{t_i} = n_{t_i}$. In the absence of pool-level covariates, all pools at a given time have the same probability of testing positive because both m_{t_i} and p_{t_i} are the same for all pools. The sampling model is

$$
\begin{aligned}
Y_{t_i}|\pi_{t_i} &\overset{\text{iid}}{\sim} \text{Binomial}(k_{t_i}, \pi_{t_i}) \\
\pi_{t_i} &= 1 - (1 - p_{t_i})^{m_{t_i}},
\end{aligned} \tag{5}
$$

where $Y_{t_i} \in \{0, 1, \ldots, k_{t_i}\}$ represents the number of pools that test positive at t_i. ~~Synthetic~~ Simulated examples follow this framework. This model is not appropriate when pools are different sizes, as is the case when the pool size does not divide the number of individual samples.

3.3.4   Efficient general pooling

The total number of individual samples may not be perfectly divisible by the chosen pool size. For example, if as many samples as possible are collected rather than a predetermined number, it is unlikely that m divides n. Suppose we test pools no larger than m^*. We can use the idealized model when m^* divides n_{t_i}, and the general model otherwise. However, when k is on the order of tens or hundreds it is computationally advantageous to model pool structure more deliberately. When m^* does not divide n_{t_i}, test $k_{t_i} - 1$ pools of size m^* and one pool of size $(n_{t_i} \bmod m^*) = m_{t_i}$. This motivates the final sampling model:

$$
\begin{aligned}
Y_{t_i,1}|\pi_{t_i,1} &\overset{\text{iid}}{\sim} \text{Binomial}(k_{t_i} - 1, \pi_{t_i,1}) \\
\pi_{t_i,1} &= 1 - (1 - p_{t_i})^{m^*};
\end{aligned} \tag{6}
$$

$$Y_{t_i,2}|\pi_{t_i,2} \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_{t_i,2})$$
$$\pi_{t_i,2} = 1 - (1 - p_{t_i})^{m_{t_i}}. \tag{7}$$

Here, the response $Y_t = (Y_{t,1}, Y_{t,2})^\intercal$ is a multivariate random vector comprising the number of positive pools among those of size m^* and a binary variable indicating the test result of the final, odd-sized pool. Because the number of individual samples may vary over time in this scenario, it may be the case that $n_{t_i} < m^*$ or $k_{t_i} \times m^* = n_{t_i}$. This model handles both situations naturally. In the former case, we have $Y_{t_i,1} \overset{\text{iid}}{\sim}$ Binomial$(0, \pi_{t_i,1})$ and in the latter, $Y_{t_i,2} \overset{\text{iid}}{\sim}$ Bernoulli(0), which imply $Pr(Y_{t_i,1} = 0) = Pr(Y_{t_i,2} = 0) = 1$; in effect, either the ideal or general model is used automatically. Obviously, $y_{t_i,1}$ and $y_{t_i,2}$ must be coded as zeros in the respective situations. This formulation can reduce the number of likelihood terms by orders of magnitude, lending a substantial speed-up. We use this model in the Notre Dame COVID-19 testing example, in which more than 80,000 data are analyzed.

3.4 Hierarchical pooling

Let $t = \{t_1, t_2, \ldots, t_n\}$ be an index set of observation times in the observation interval $\mathcal{T}$ with observed responses $y_t = \{y_{t_1}, \ldots, y_{t_n}\}$ where $y_{t_i} \in \{1, \ldots, k_{t_i}\}$, corresponding to the ideal scenario described above. Additionally, let W_t be the set of latent prevalence values at times t. The hierarchical model is

$$Y_{t_i}|\pi_{t_i} \overset{\text{iid}}{\sim} \text{Binomial}(k_{t_i}, \pi_{t_i})$$
$$\pi_{t_i} = 1 - (1 - p_{t_i})^{m_{t_i}}$$
$$p_{t_i} = \phi(W_{t_i} + \mu)$$
$$W_t|C \sim \mathcal{GP}(\mathbf{0}, C_\theta) \tag{8}$$
$$\mu \sim p(\mu)$$
$$\theta \sim p(\theta).$$

Replace the first two lines with the appropriate sampling model in non-ideal scenarios. In the absence of covariates, we model a scalar mean μ. Instead modeling the mean as a linear combination of predictors $\mu = X\beta$ would be a natural extension. Furthermore, time-varying coefficients could also be used where $\mu_{t_i} = X_{t_i}\beta$.

~~The model used to infer prevalence from individual samples is the same as (8) with the prevalence transformation omitted and the sampling model~~

$$z_{t_i}|p_{t_i} \overset{\text{iid}}{\sim} \text{Binomial}(k_{t_i}, p_{t_i}),$$

~~where k_{t_i} is now the number of individuals tested at t_i and $z_{t_i} \in \{1, \ldots, k_{t_i}\}$ is the observed number of infected individuals.~~

3.5 Hierarchical data integration

Our data integration procedure requires combining pooled information with individual samples. The individual samples contain additional information pertinent to ~~small area~~ estimation, such as a specific species with the Congo Basin bat study. For this framework, we build off the notation from Equation 8.

Let $z_{t_i,j}$ be the number of infected individuals at t_i for the jth subpopulation. Similarly, $k_{t_i,j}$ is the

number of individuals tested from the jth subpopulation at time t_i.

$$
\begin{aligned}
Z_{t_i,j}|p_{t_i,j} &\overset{\text{iid}}{\sim} \text{Binomial}(k_{t_i,j}, p_{t_i,j}) \\
Y_{t_i}|\pi_{t_i} &\overset{\text{iid}}{\sim} \text{Binomial}(k_{t_i}, \pi_{t_i}) \\
\pi_{t_i} &= 1 - (1 - p_{t_i})^{m_{t_i}} \\
p_{t_i} &= \phi(W_{t_i} + \mu) \\
p_{t_i,j} &= \phi(W_{t_i} + \mu_j) \\
W_t|C &\sim \mathcal{GP}(\mathbf{0}, C_\theta) \\
\mu_j|\mu &\overset{\text{iid}}{\sim} N(\mu, \sigma_\mu^2) \\
\mu &\sim p(\mu) \\
\theta &\sim p(\theta) \\
\sigma_\mu^2 &\sim p(\sigma_\mu^2).
\end{aligned}
\tag{9}
$$

where μ_j is the subpopulation specific offset. `It is important to note that this framework does assume that all of the subpopulations share the same overall trend, but the trend can be shifted up or down with the` μ_j `values. If this is not reasonable, hierarchical structure would be needed in the` W_t `processes.`

3.6 Priors and computation

In the interest of propagating uncertainty as thoroughly as possible, we perform fully-Bayesian inference on all model parameters and hyperparameters by use of Markov chain Monte Carlo (MCMC). Probabilistic inference on GP hyperparameters is challenging in this setting due to the posterior dependence between latent variables and hyperparameters; the coupling can be resolved by intensive sampling, but the $\mathcal{O}(N^3)$ scaling behavior of GP prediction renders such intensive, repeated computation impractical. The data sets considered here are sufficiently small that inference in Stan [22] remains practical. `For all of the analyses, four Markov chains were used. Generally, 2000 - 3000 iterations with 1000 - 1500 iterations for warmup were sufficient for parameters to converge. The analyses were conducted on a 3.1 GHz Dual-Core Intel Core i7 took less than about 20 minutes when running 4 chains in parallel.` For larger data sets, the dependence structure must be addressed directly. Several methods exist; see [23] for a review.

We use normal and half-normal priors on μ and σ, respectively. Values of ℓ less than the shortest time between observations (<u>low</u>) or greater than the total observation interval (<u>high</u>) are not identified. Accordingly, an inverse-gamma prior `is placed` on ℓ to inform the range of plausible values between <u>low</u> and <u>high</u>.

4 `Simulation` studies

Dynamic prevalence estimation from pooled samples is first evaluated through two studies of simulated data. The ~~synthetic~~ `simulated` studies are used to evaluate the proposed method's ability to recover known, underlying prevalence in settings similar to those of the case studies. The ~~synthetic~~ `simulated` studies are designed to reflect characteristics present in the data used in the applications. ~~Clearly,~~ It is not possible to simulate a perfect match to the unobserved prevalence processes that gave rise to the observed data, but we approximately match the order of magnitude, overall trend, and sampling design. In the first study, true prevalence ranges from 0 to 0.12 and oscillates slowly in that range. The prevalence is 'observed' at 25 evenly spaced times within a 1000 day interval. 45 individual test results are simulated for each observed prevalence and 15 pools of size 3 ($k = 15$, $m = 3$) are constructed. In the second study, true prevalence does not exceed 0.05 and tests are simulated every day for 150 days. 500 individual test results are simulated each day and pool size $m = 10$ is used, so 50 pools are required each day ($k = 50$, $m = 10$). Here, a larger pool size is tolerated because prevalence is low. This would not be known a priori, but may be available through domain knowledge, past analyses, or initial testing. In both studies, $m = 1$ is informed by $k \times m$ individual data, $m = 1^*$ is informed by k individual data, and $m > 1$ are informed by k pools of size m.
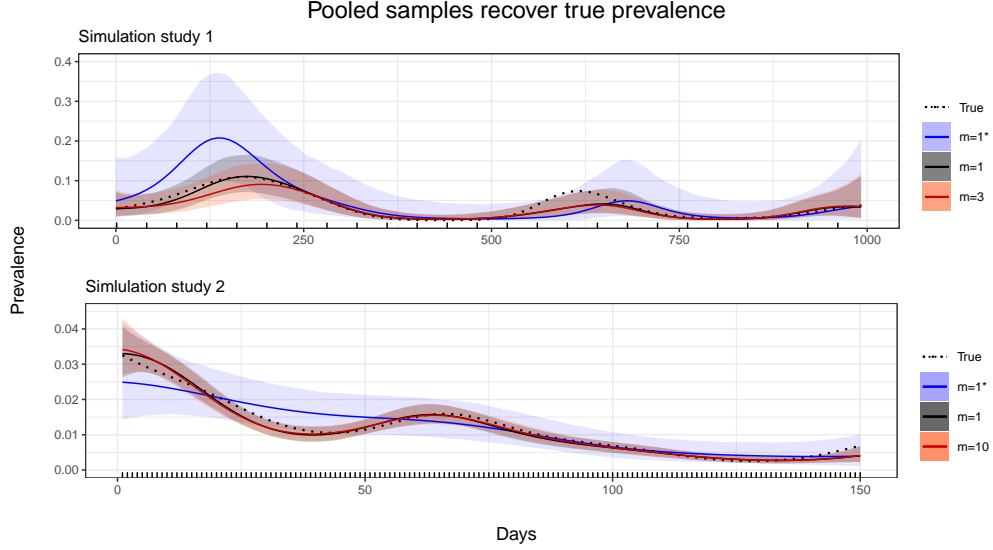
Figure 1: Estimated ~~median~~ mean curves and 95% credible regions are displayed for each ~~synthetic~~ simulated study under each sampling strategy. Rug marks indicate when sampling occurred. In both cases, the pooled estimates closely track true, unobserved prevalence (black, dotted) and $m = 1$ estimates (black, solid). Estimates from limited individual sampling (blue, solid) are less precise and fail to accurately recover true prevalence.

Table 1: GP hyperparameter estimates for ~~synthetic~~ simulated studies (**1** and **2**). Estimated posterior means and 95% CIs are reported for each parameter and sampling regime.

1	True	$m = 1$	$m = 3$	$m = 1^*$
ℓ	100	89.7 (53.8, 136.9)	91.3 (55.6, 139.1)	92.8 (50.0, 164.6)
σ	0.5	0.59 (0.31, 1.00)	0.56 (0.29, 0.98)	0.67 (0.32, 1.18)
μ	-2	-1.99 (-2.46, -1.50)	-2.00 (-2.46, -1.53)	-1.90 (-2.50, -1.38)
2	True	$m = 1$	$m = 10$	$m = 1^*$
ℓ	25	23.3 (16.8, 32.4)	24.4 (17.6, 34.4)	34.8 (18.9 64.7)
σ	0.25	0.38 (0.20, 0.71)	0.39 (0.21, 0.72)	0.36 (0.15, 0.74)
μ	-2.33	-2.25 (-2.63, -1.80)	-2.24 (-2.63, -1.80)	-2.25 (-2.68, -1.79)

In both simulated studies, there is strong agreement between estimated curves for individual ($m = 1$) and pooled ($m = (3, 5)$) data, displayed in Figure 1. In the top panel, pooled and individual curves track the true prevalence closely with a clear but slight exception around day 625. The budgeted individual curve ($m = 1^*$) is both less precise and less accurate in general. The second study (Figure 1, panel 2), in which much lower prevalence is estimated from more data over a shorter interval, provides even more compelling results. True prevalence is estimated well by individual and pooled data, and the two are largely indistinguishable. Furthermore, using budgeted individual data grossly over-smooths the curve, reducing the process to an approximately linear function. Notably, where true prevalence is 0, pooled results are disproportionately informative because it is known that exactly $k \times m$ individuals are negative, whereas individual results at the same testing budget indicate only that k individuals are negative—pooling lends greater precision through sample size when prevalence is near zero. Table 1 supports these results; compared to $m = 1$ and pooled estimates, $m = 1^*$ estimates are either less accurate, less precise, or both, with the exception of μ in the second ~~synthetic~~ simulation study. Together, these simulation studies demonstrate that 1) true, underlying prevalence can be estimated using pooled sample data together with the proposed hierarchical model, 2) for a fixed number of tests, use of pooled sample data confers substantial gains in precision and accuracy, and 3) results from pooled data are consistent with results from individual data, which require m times as many

tests and therefore m times higher variable cost. ~~Having established that true prevalence can be efficiently estimated from pooled data, we proceed with analyzing the data described in Section 2.~~
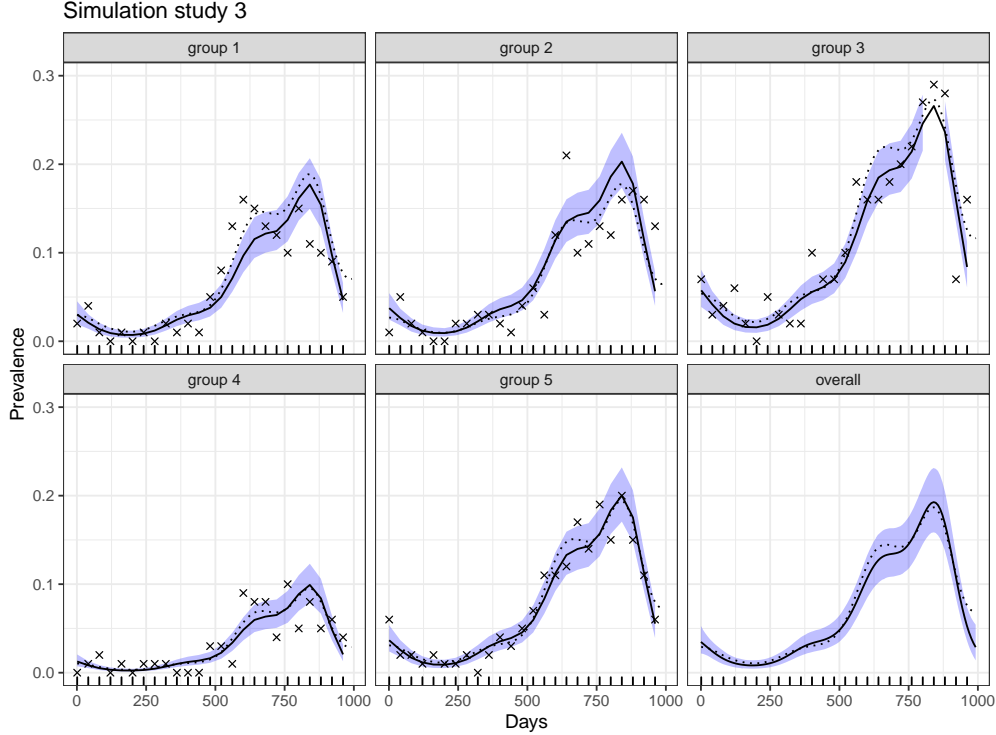
Simulation study 3

Figure 2: This figure shows the prevalance dynamics for each of the five groups as well as the overall mean prevalence. Estimated ~~median~~ mean curves are dotted lines, true latent prevalences are solid lines, and 95% credible regions are displayed with blue shading. The proportion of observed positive individuals are shown with the x's. The estimated dynamics for both the small area estimation problem and the overall prevalence closely track the true prevalence.

The third ~~synthetic~~ simulated data study explores the data-integration procedure to combine pooled and individual data. This simulation ~~small area estimation problem~~ parallels the setting in ~~synthetic~~ simulated study 1. The true prevalence is generally less than about 0.25, and considerably less for some of the subpopulations. The prevalence is 'observed' at 25 evenly spaced times within a 1000 day interval. For the pooled results 45 individual test results are simulated for each observed prevalence and 15 pools of size 3 ($k = 15$, $m = 3$) are constructed. ~~Individuals are a random sample from the population the metadata associated with those individuals are not available.~~ Additionally, 500 individual test results are simulated each day, corresponding to 100 for each of 5 separate groups. Each group will have a different prevalence driven by the different μ_i values from Equation 9, which shift the latent prevalence curve and result in shared, but shifted prevalence curves across the groups.

The data integration procedure enables estimation of the dynamics of both the overall prevalence as well as the group-specific prevalence. In particular, the estimated curves (dotted lines) in Figure 2 closely track the patterns from the observed data. Our procedure results in more precise estimates of the group dynamics that could be estimated from the individual samples alone by sharing information with pooled samples in a hierarchical framework. This sharing also enables estimation of prevalence dynamics of sparsely surveyed subpopulations~~, or in other words small area estimation of these dynamics~~. Similarly, using the individual samples provides more precise estimates of the overall dynamics than could be estimated from just the pooled samples. In addition to the dynamics curves, the credible intervals from the models contain the true estimated prevalence. Having verified the efficacy of this procedure, it will be applied to a setting with Congo Basin bats.

5 Results

5.1 Case study 1: Congo Basin bats

Pooled estimates ($m = 3$ and $m = 5$), representing 3- and 5-fold reductions in the number of tests, closely match the $m = 1$ curve obtained through universal individual testing, both in ~~median~~ mean and uncertainty intervals (Figure 3). The $m = 1^*$ curve, which has equal testing cost to $m = 3$ and greater cost than $m = 5$, is both less precise and a poorer fit to the best alternative, $m = 1$. Table 2 tells a similar story. Parameter estimates from pooled data closely match $m = 1$ estimates in terms of both mean and 95% interval endpoints. Estimates from $m = 1^*$ are either biased (from $m = 1$) or less precise.
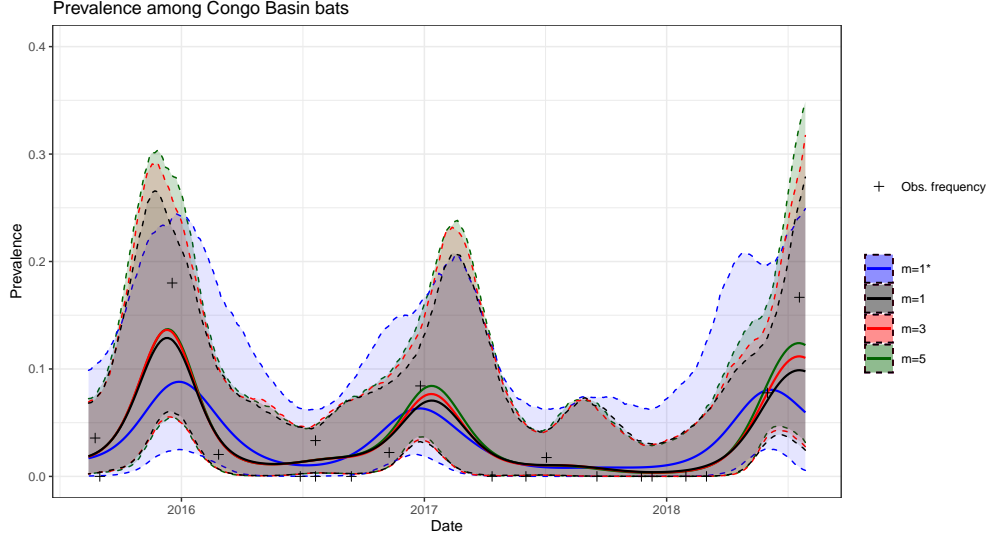

Figure 3: Prevalence among Congo Basin bats considered in this study is estimated in four ways; ~~medians~~ means and 95% quantile-based credible regions are displayed for each. Universal individual testing estimates (black) and pooled estimates (red, green) exhibit high visual similarity; pooling introduces slight, additional uncertainty. Estimates from limited individual testing (blue) capture the same overall trend but deviate from other estimates and possess more uncertainty throughout.

Table 2: GP hyperparameter estimates under each sampling regime for case study 1. Estimated posterior means and 95% CIs are reported for each parameter.

	$m = 1$	$m = 3$	$m = 5$	$m = 1^*$
ℓ	77.3 (45.3, 132.5)	75.9 (45.2, 122.7)	76.4 (44.5, 127.3)	89.1 (48.9, 167.8)
σ	0.63 (0.30, 1.08)	0.65 (0.31, 1.12)	0.67 (0.34, 1.12)	0.56 (0.05, 1.12)
μ	-1.95 (-2.47, -1.48)	-1.94 (-2.47, -1.48)	-1.93 (-2.46, -1.43)	-1.93 (-2.56, -1.43)

5.2 Case study 2: Congo Basin bats data integration

Combining pooled data, without species information, with individual data enables efficient estimates of the prevalence dynamics for species. The species estimates also inform the overall mean, which is jointly estimated from the pooled data and, hierarchically, through the individual species means. In disease ecology, survey efforts of hard to reach populations or rare species may ~~not~~ result in ~~large~~small sample sizes or sparse temporal coverage. This can be seen in Figure 4 where many species have large gaps in ~~in~~ survey effort over time; however, our data integration procedure enables estimates across those gaps by borrowing

information from the overall dynamics which are informed by both other species and pooled samples through the hierarchical structure of the overall mean prevalence across groups.

There was an approximately 9 month window in early 2013 without any survey effort - either of individual or pooled samples. As we'd hope the uncertainty across all species and the overall mean increases during this window. In early 2014 we see an uptick in prevalence across many of the species that is also reflected in the pooled samples. The overall prevalence is not as high as simulation study 3, but Microteropus pusillus, Epomops franqueti, Megaloglossus woermanni, and the overall groups all have a clear increase in early 2014. While Myonycteris torquota and Pipistrellus sp. do not have obvious increases at this same time, less sampling effort is given for those species; nevertheless, the mean prevalence stays near zero during this time period, but uncertainty intervals do increase based on information from other species. This behavior suggests that our approach to combining inferences across species is reasonable as the dynamics appear similar.
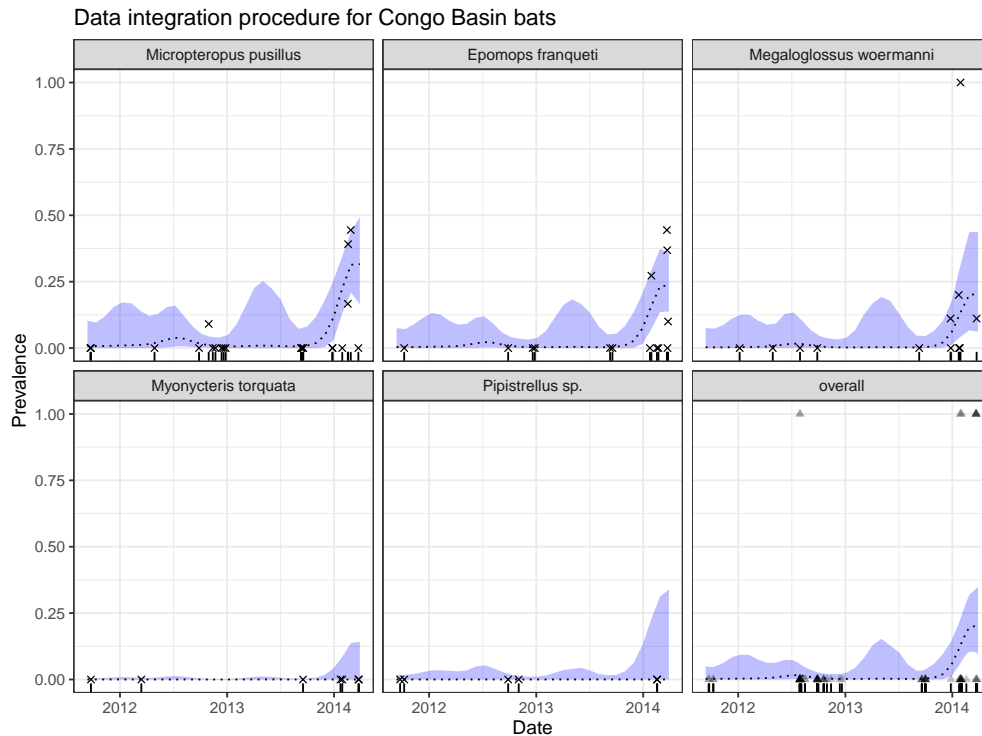

Figure 4: Prevalence among Congo Basin bats is estimated using individual data from 5 species as well as pooled data. ~~Medians~~ Means and 95% quantile-based credible regions are displayed for each species as well as the overall mean. The rug marks indicate survey dates, the stars for individual species correspond to proportion of positive samples during that time period, and the triangles indicate whether a pool was positive or negative. It is important to note that some of the highest individual prevalence points are driven by small sample sizes.

5.3   Case study 3: Asymptomatic testing at Notre Dame

Across all testing strategies, the number and frequency of tests lends greater precision, compared to the previous example. Figure 5 displays estimated curves from $m = 5$ and $m = 10$, which are nearly indistinguishable from the universal individual testing curve, $m = 1$. Due to the low prevalence and testing intensity, pooled testing is extremely efficient—an order-of-magnitude reduction in testing costs affects estimates almost imperceptibly. However, uncertainty intervals for $m = 1^*$ are again far wider and the estimates are clearly over-smoothed. Table 3 also indicates that pooled estimates recover $m = 1$ estimates with high fidelity, but $m = 1^*$ estimates differ considerably in expectation and precision.

Observed frequencies at most times are consistent with the posterior predictive distribution, but very low observed frequencies coupled with large sample sizes are represented only in the extreme left tails of predictive distributions at the respective times. Incorporating demographic information or correlation among individuals tested at the same time may resolve this.
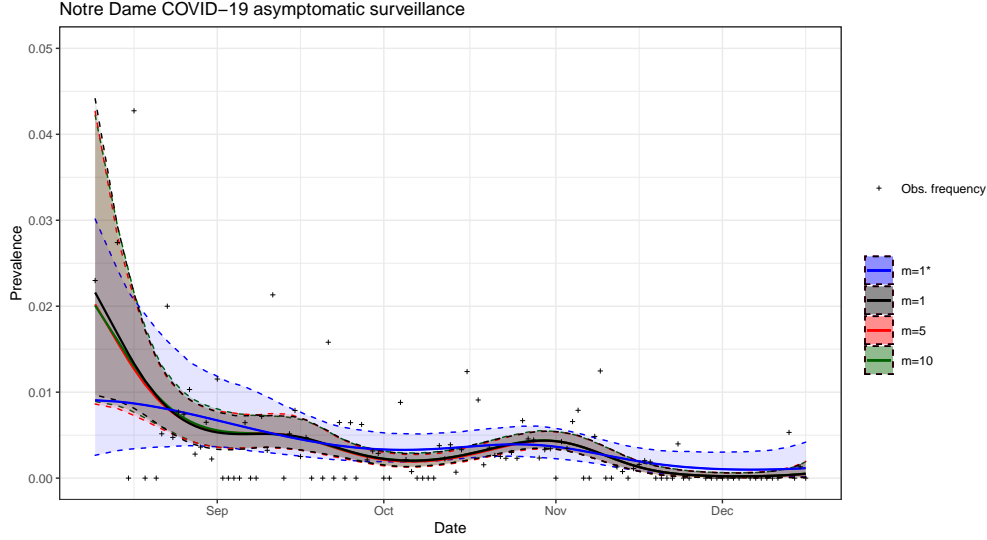

Figure 5: Four estimated ~~median~~ mean prevalence curves and associated 95% credible regions of prevalence among asymptomatic individuals at Notre Dame are displayed. For most time points, universal individual (black) and pooled (red, green) estimates are visually indistinguishable. Estimates from limited individual testing (blue) are, again, deviant and more uncertain, on average.

Table 3: GP hyperparameter estimates under each sampling regime for case study 2. Estimated posterior means and 95% CIs are reported for each parameter.

	$m = 1$	$m = 5$	$m = 10$	$m = 1^*$
ℓ	19.7 (13.4, 27.8)	20.2 (13.5, 28.3)	20.3 (13.8, 27.8)	27.4 (16.3, 47.4)
σ	0.60 (0.32, 1.08)	0.60 (0.32, 1.05)	0.58 (0.31, 1.03)	0.39 (0.10, 0.88)
μ	-2.49 (-3.02, -1.84)	-2.49 (-3.00, -1.81)	-2.50 (-3.01, -1.80)	-2.57 (-3.01, -2.02)

5.4 Sampling variability

Reported results from the Congo Basin bat and Notre Dame COVID-19 surveillance studies are based on sampled pool configurations and subsampled individual results. Other configurations and subsamples are possible, and the sampling variation introduced by this design is not formally represented in the reported results. The Notre Dame study results entail less sampling variability because the sample size is two orders of magnitude larger. We evaluate the impact of the design on the Congo Basin bat example through resampling. 25 pooled ($m = 3$) and 25 ~~susbsampled~~ subsampled individual data sets are resampled from the complete bat data set and curves are fit to each. ~~Median~~ Mean curves, along with the original $m = 1$ curve and its 95% credible interval, are reported in Figure 6; this figure illustrates variation induced by the pooled sampling assignment.

In practice, pooling is not achieved by subsampling individual results. Instead, pooled testing either incorporates information from more individuals than would otherwise be tested or tests all individuals at a lower cost. Figure 6 is included for completeness and suggests that the pooled results displayed in Figure 3 are typical under the pool sampling design. Individual results at the same budget, however, are far more variable. Consider a scenario where $k \times m$ individuals could be sampled, but only k tests are budgeted.
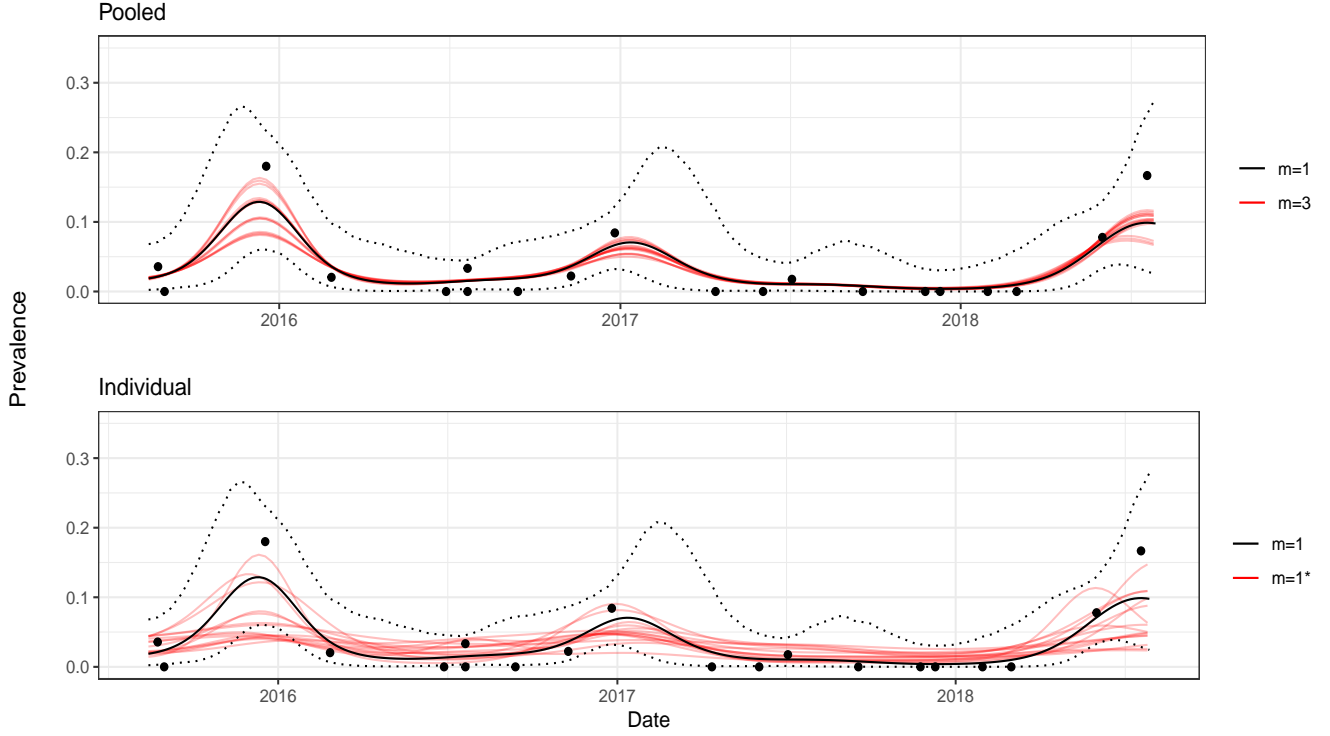
13

Figure 6: Estimated ~~median~~ mean curves for 25 resampled pool configurations and 25 subsampled individual configurations are displayed over the estimated ~~median~~ mean from all available date and corresponding 95% credible region (identical to Figure 3, $m = 1$). Pooled estimates are relatively stable under reconfiguration, but estimates from subsampled individual data are quite sensitive to the subsampling procedure.

Pooled testing permits the incorporation of information from all $k \times m$ individuals and subsequent analyses provide stable approximations to estimates generated from universal individual testing data.

5.5   Conclusion

The simulation studies have established that pooled testing data together with latent GP regression can be used to estimate prevalence over time and, importantly, data integration procedures can be used ~~for small area estimation~~ to estimate prevalence within specified subpopulations. The case studies further established that pooled estimates closely match the far more costly alternatives ($m = 1$) and that pooled estimates are more precise than budgeted alternative ($m = 1^*$). Additionally, combined pooled samples with individual samples can share information from pools with metadata associated with individual samples for more precise estimation. The realized pool assignments and resulting estimates are typical under the pool simulation scheme and not particularly sensitive to pool assignments.

Smoothed and interpolated prevalence estimates are more representative of prevalence dynamics than sets of pointwise estimates. Obtaining such estimates requires more intensive sampling and careful inference. The inferential efficiency of pooled testing, which had previously been shown only in static settings, is seen to apply to dynamic modeling as proposed in this article also.

6   Discussion

Population prevalence as a function of time is estimable from pooled samples. Compared to individual testing, pooling and data integration serves to greatly reduce testing costs without qualitatively affecting estimates

or uncertainty; at a fixed budget, pooling generates more precise estimates ~~and can be used for small area estimation~~. Applied to surveillance of reservoir host populations, the proposed methodology enables efficient, precise estimation of pathogen prevalence dynamics.

Throughout, we assumed that unlimited, universal testing best approximates true prevalence, among the strategies considered. A two-stage design is an interesting intermediary between pooled testing and individual testing. It would likely be more costly than pooling and less accurate than individual testing but may achieve a more favorable cost/accuracy trade-off than either. We leave this for future work.

Other natural extensions to this work exist, some of which immediately resolve current limitations. These extensions belong to three thematic categories: incorporating additional information, using estimates to inform sampling, and elaborating on the chosen inference tools.

Known values of sensitivity and specificity may additionally be incorporated at a commensurate expense to precision. A limitation of pooling is the potential for dilution-induced changes in sensitivity and specificity as a function of pool size. Resolving this entails a trade-off between cost and precision. Similarly, modeling covariance jointly in space and time would be an exciting extension applicable to sufficiently rich data sets.

In online learning settings, where data are analyzed as they become available, various forms of adaptive sampling may be used to design optimal pool sizes at subsequent times. It may also be advantageous to tune sampling frequency in real-time to ensure that dynamics of interest are identified. These adaptive sampling strategies would ensure testing resources are efficiently and effectively expended.

Lastly, the nonparametric tools described in this article were used with relatively basic specifications to establish the general applicability. The model could be strengthened by prior-encoded domain knowledge, alternative covariance functions, and covariate-informed mean structure, depending on the application. One other specific extension would be a process to enable different temporal trends for subgroups, rather than subgroups sharing the same trend with a constant shift. This could be done by manipulating the W_t processes from Equation 8. Furthermore, as the model is expanded to handle more information, efficient or approximate computation will become increasingly relevant. We look forward to these extensions as exciting future work.

Author Contributions

Blinded for review

Data Availability

Data and code used in this work are available but ~~at~~ blinded for review.

References

[1] Raina K Plowright, Colin R Parrish, Hamish McCallum, Peter J Hudson, Albert I Ko, Andrea L Graham, and James O Lloyd-Smith. Pathways to zoonotic spillover. Nature Reviews Microbiology, 15(8):502–510, 2017.

[2] Daniel Becker, Peggy Eby, Wyatt Madden, Alison Peel, and Raina Plowright. Ecological conditions experienced by bat reservoir hosts predict the intensity of hendra virus excretion over space and time. bioRxiv, 2021.

[3] Thijs Kuiken, Frederick A Leighton, Ron AM Fouchier, James W LeDuc, Joseph S Malik Peiris, Alejandro Schudel, Klaus Stöhr, and ADME Osterhaus. Pathogen surveillance in animals. Science, 309(5741):1680–1681, 2005.

[4] Raina K Plowright, Daniel J Becker, Hamish McCallum, and Kezia R Manlove. Sampling to elucidate the dynamics of infections in reservoir hosts. Philosophical Transactions of the Royal Society B, 374(1782):20180336, 2019.

[5] Christian Gollier and Olivier Gossner. Group testing against covid-19. Technical report, EconPol Policy Brief, 2020.

[6] Jaya Garg, Vikramjeet Singh, Pranshu Pandey, Ashish Verma, Manodeep Sen, Anupam Das, and Jyotsna Agarwal. Evaluation of sample pooling for diagnosis of COVID-19 by real time-PCR: A resource-saving combat strategy. Journal of medical virology, 93(3):1526–1531, 2021.

[7] Andrew Hoegh, Alison Peel, Wyatt Madden, Manuel Ruiz-Aravena, Aaron Morris, Alex Washburne, and Raina Plowright. Estimating viral prevalence with data integration for adaptive two-phase pooled sampling. 2021.

[8] Zitao Liu, Lei Wu, and Milos Hauskrecht. Modeling clinical time series using gaussian process sequences. In Proceedings of the 2013 SIAM International Conference on Data Mining, pages 623–631. SIAM, 2013.

[9] Elizabeth Buckingham-Jeffery, Valerie Isham, and Thomas House. Gaussian process approximations for fast inference from infectious disease data. Mathematical biosciences, 301:111–120, 2018.

[10] Hirotugu Akaike. Fitting autoregressive models for prediction. Annals of the institute of Statistical Mathematics, 21(1):243–247, 1969.

[11] Malay Ghosh and JNK Rao. Small area estimation: an appraisal. Statistical science, 9(1):55–76, 1994.

[12] Danny Pfeffermann. New important developments in small area estimation. Statistical Science, 28(1):40–68, 2013.

[13] JNK Rao and Mingyu Yu. Small-area estimation by combining time-series and cross-sectional data. Canadian Journal of Statistics, 22(4):511–528, 1994.

[14] Isabel Molina, Balgobin Nandram, and JNK Rao. Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. The Annals of Applied Statistics, 8(2):852–885, 2014.

[15] Gauri Sankar Datta, Malay Ghosh, R Steorts, and J Maples. Bayesian benchmarking with applications to small area estimation. Test, 20(3):574–588, 2011.

[16] Charles Kumakamba, Fabien R Niama, Francisca Muyembe, Jean-Vivien Mombouli, Placide Mbala Kingebeni, Rock Aime Nina, Ipos Ngay Lukusa, Gerard Bounga, Frida N'Kawa, Cynthia Goma Nkoua, et al. Coronavirus surveillance in wildlife from two Congo basin countries detects RNA of multiple species circulating in bats and rodents. PloS one, 16(6):e0236971, 2021.

[17] Mark D Fox, David C Bailey, Michael D Seamon, and Marie Lynn Miranda. Response to a COVID-19 outbreak on a university campus—Indiana, August 2020. Morbidity and Mortality Weekly Report, 70(4):118, 2021.

[18] University of Notre Dame. Covid-19 protocols and policies: Fall 2020 data, 2021.

[19] Silvestre Colón, Ganapati P Patil, and Charles Taillie. Estimating prevalence using composites. Environmental and Ecological Statistics, 8(3):213–236, 2001.

[20] William H Swallow. Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology (USA), 1985.

[21] Marie Reilly. Optimal sampling strategies for two-stage studies. American journal of epidemiology, 143(1):92–100, 1996.

[22] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. Journal of statistical software, 76(1):1–32, 2017.

[23] Maurizio Filippone, Mingjun Zhong, and Mark Girolami. A comparative evaluation of stochastic-based inference methods for Gaussian process models. Machine Learning, 93(1):93–114, 2013.