Catchy Title Goes Here

Laura Sampson[1] and Matthew Ferrari[1]

[1] *Center for Infectious Disease Dynamics, Pennsylvania State University, State College, PA, 16801*

I. INTRODUCTION

Dramatic improvements in the coverage of vaccination with measles containing vaccines have lead to significant reductions in the incidence of measles disease and thea associated childhood mortality [REF - Dabbagh et al]. However, despite these improvements at the national scale, significant heterogeneity in both vaccination coverage and disease incidence remains at the sub-national level [REF Takahashi]. Given the high transmission rate of measles, the proportion of the population that must be immunized to achieve and maintain elimination of measles is likely to be greater than 90% in most populations [REF - Goodson?] (the exact level will be modulated by locally specific conditions just as popualtion density and contact rates).

The size and age-distribution of the population susceptible to measles (hereafter we refer to this as the "susceptible persons") depends on the competing prossess of susceptible recruitment through birth and susceptible loss through either natural infection or vaccination. Quantifying the size and age distribution of susceptible persons to measles is a critical tool in evaluating outbreak risk, the performance of vaccination programs, and developing vaccine-based interventions. The WHO recommends that programs monitor the accumulation susceptible persons using "good quality data" for all countries and that these should be made at the sub-national level for large countries and those close to measles elimination [REF – Wkly Epid Rec].

Direct observation of the distribution of susceptible persons is challenging as it requires a sero-survey which can be cost-prohibitive (Winter et al JID paper). A sero-survey can provide a well resolved cross-sectional estimate of the distribution of suscpetible persons, it cannot, by itself, quantify the relative contribution of natural and vaccine derived immunity as current serological diagnostics cannot distinguish between these two sources [REF - Winter JID]; however, the relative contribution of these sources can be estimated through modeling [REF- Winter JID]. In the absence of a sero-survey, the number and age-distribution of susceptible persons can be estimated though demographic models that account for inputs through births and immunization via both vaccination an natural infection [Winter et al JID, Winter et al Madagascar, Merler et al http://www.thelancet.com/pdfs/journals/laninf/PIIS1473-3099(17)30421-8.pdf]. In this approach, the contribution of natural immunity is modeled as the result of a dynamic transmission model [REF] which requires explicit assumptions about the age-sepcific force of infection via a Who Acquires Infection From Whom matrix that describes the age-specific rate of infectious contacts from each age class to each other. Direct estimation of this WAIFW matrix is challenging from disease incidence or seroprevalence data [though see Whittaker and Farrington etc.], and thus it is common to make simplfying assumptions that either the WAIFW matrix is constant for all ages (i.e. a uniformly well-mixed system) or that the WAIFW matrix is a scalar function of some other measureable interaction process, such as contacts measures via diary studies [REF – Mossong POLYMOD study].

The catalytic model was initially proposed as a method for estimating the age-specific force of infection from cross-sectional age-specfic sero-prevalence data [REF - Griffiths]. This model, which represents the probability of immunization via natural infection at specific age a as the cumulative sum of age-specific force of infection prior to age a, was then adapted by Grenfell and Anderson [REF - 1985] for use with cross-sectional, age-specific case observations. The resulting fitted model can then be used to estimate the age-specific distribution of susceptible persons. Thus, in the absence of a sero-survey it is possible to estimate seroprevalence from routinely collected age-specific case reporting.

Original applications of the catalytic model assumed that natural infection was the only source of immunity [REF - Griffiths]. In the vaccine era, however, the age distribution of susceptibility is generated by the sum of the rates of vaccination and natural infection. Though measles vaccination is recommended to follow a specific schedule, with children receiving two doses at prescribed ages, in practice, the ages at which receive an immunizing dose of measles containing vaccine (MCV) may be highly variable because of variaion in access to vaccination services [REF Metcalf; Takahashi; others] and the multiple vaccination initiatives that are employed: e.g. routine vaccination, supplementary immunization activities, child and maternal health days, school-based immunization drives, etc. Recent work has highlighted significant variability both in the maximum vaccination coverage achieved and the timeliness of vaccination, that is, the proportion of children receiving vaccination at any given age. Thus, vaccination can itself be modeled as an age-specific hazard rate that accounts for the sum of multiple forces. Failing to account for the age-specific pattern of vaccination may lead to a biased interpretation of case-data, as children who might be vaccinated later than the recommended age remain susceptible may contribute to incident cases.

Further, vaccination does not necessarily imply immunization. The efficacy of measles vaccine is often assumed

to be ... (see Uzicanin article) though is also known to improve with age as older children are more likely to have lost maternally transferred antibodies (se many). The effectiveness of vaccine delivered in field settings may also vary dramatically due to stability and effectiveness of the vaccine cold chain (Doshi et al 2017 Vaccine). A comparison of the age-distribution of vaccination and seroprevalence may highlight areas with low effectiveness, and absent a sero-survey this assessment could be made using age-specific case records.

Here we present an extension of the catalytic model that represents the age-specific proportion immune as the cumulative sum of the hazards of both natural infection (force of infection) and effective immunization via vaccination. We illustrate how this model can be used to estimate the age-specific sero-prevalence, and also the age-specific rates of vaccination and force of infection, and vaccine effectiveness using age-specific vaccine coverage data and case-records. We illustrate performance of this model using both simulated data, and measles case surveillance data from DRC combined with vaccination coverage surveys conducted as part of the 2013-14 DHS. A contemporary measles sero-survey conducted during the 2013-14 DHS allows us to validate the performance of our estimates of sero-prevelence against direct measurements. We finally present a fit of the model to the surveillance data, vaccine coverage data, and sero-survey and discuss opportunities for combining data sources.

II. METHODS

A. Competing Rates Model

The classic catalytic model for disease infection, developed in the late 1950's, gives the probability of immunity at age a as

$$p(\text{immune}|a) = 1 - \exp\left(-\int_0^a f(a')da'\right), \qquad (1)$$

where $f(a)$ is the *force of infection* at age a, which can be thought of as the rate of infection at a particular age. This expression is valid in the absence of vaccination, as in this case infection is the only source of immunity. In the case of measles, this expression gives the probability of an individual at age a being seropositive for measles.

In situations in which vaccination is present, vaccination provides a second means of acquiring immunity - the 'force of vaccination,' or 'vaccination hazard' (vh). If we represent this as $v(a)$, then we can extend Eq. 1 to give the probability of an individual testing seropositive at age a as

$$\begin{aligned} p(\text{immune}|a) &= 1 - \exp\left(-\int_0^a f(a') + v(a')da'\right) \\ &= 1 - \exp\left(-\int_0^a f(a') - \int_0^a v(a')da'\right). \end{aligned} \qquad (2)$$

The functional forms of $f(a)$ and $v(a)$ are free to be specified. In this study, we choose to use un-normalized Weibull distributions to parameterize both of these functions, as they have been shown to be sufficiently flexible to match a range of possible forces of infection and vaccination hazards[REF]. Thus $f(a) \to f(a|\psi)$ and $v(a) \to v(a|\theta)$, where ψ and θ are vectors of the parameters we use for the Weibull distribution - height (η), scale (α), and shape (β). This gives six parameters that fully specify the forms of $f(a)$ and $v(a)$ from Eq. 2 - α, β, and η for two independent Weibull distributions. To allow for the fact that not all vaccinations produce immunity, we introduce a seventh parameter - the vaccine effectiveness (γ). This enters the equation as a multiplier on $v(a)$ that ranges between 0.0 and 1.0.

Besides the flexibility of the distribution, another appealing aspect of the Weibull as our choice of parameterization is that it can be integrated analytically, as

$$\int_0^a g(a';\eta,\alpha,\beta) = \frac{\beta}{\alpha}\eta\left(1 - \exp(-(x/\alpha)^\beta)\right). \qquad (3)$$

We re-absorb the factor of β/α on the r.h.s. of this equation into the parameter η, and have a final expression for the probability of immunity at a particular age

$$\begin{aligned} p(\text{immune}|a) = 1 - \exp\big\{ &- \eta_f\left(1 - \exp(-(a/\alpha_f)^{\beta_f})\right) \\ &- \eta_v\gamma\left(1 - \exp(-(a/\alpha_v)^{\beta_v})\right)\big\}, \end{aligned} \qquad (4)$$

which is parameterized by seven total parameters - $\{\eta_f, \alpha_f, \beta_f, \gamma, \eta_v, \alpha_v, \beta_v\}$.

FIX Need to add a description of the model including a constant, and a discussion of the fact that we end up using this for the real datasets in many cases.

This expression gives the probability of observing a seropositive individual at age a, which accounts for one of our three datasets. The other two are vaccination data and case data. Using the function described above for vaccination hazard, the probability of an individual having been vaccinated by age a is given by

$$p(\text{vaccinated}|a) = 1 - \exp\left\{-\eta_v(1 - \exp\left(-(a/\alpha_v)^{\beta_v}\right)\right\}. \qquad (5)$$

Finally, the probability of an individual being recorded as a case at age a is the force of infection at that age multiplied by the probability that an individual is susceptible at that age. The probability of susceptibility is of course $1 - p(\text{immune}|a)$, and so the probability of observing a case at age a is

$$p(\text{case}|a) = \left\{1 - \exp\left[-\eta_f \left(\frac{\beta_f}{\alpha_f}\right)\left(\frac{a}{\alpha_f}\right)^{\beta_f - 1}\right] \exp\left[(a/\alpha_f)^{\beta_f}\right]\right\} \times \{1 - p(\text{immune}|a)\}\,. \quad (6)$$

The data we work with (described in detail in Sec. II E) consists of the number cases as a function of age (in years), serology tests and results as a function of age, and vaccination status as a function of age for a sample of individuals within a particular province. The likelihood for observing this dataset given values for the model parameters as

$$\begin{aligned}\log \mathcal{L}(\mathbf{c}, \mathbf{v_t}, \mathbf{v_o}, \mathbf{s_t}, \mathbf{s_o} | \eta_f, \alpha_f, \beta_f, \gamma, \eta_v, \alpha_v, \beta_v) \\ = \sum_a B\left(v_o(a), v_t(a); p(\text{vaccinated}|a)\right) \\ + \sum_a B\left(s_o(a), s_t(a); p(\text{immune}|a)\right) \\ + M\left(\mathbf{c}; p(\text{case}|a)\right), \quad (7)\end{aligned}$$

where s_t is the total number of individuals tested for IgM seropositivity, and s_o is the number of positive tests; and v_t is the number of individuals surveyed about vaccination status, while v_o gives the number who have been vaccinated. B represents a binomial probability, and M represents a multinomial distribution, and the sums are over age classes. As noted, each data point includes the age of the individual in question.

B. Model Comparison

For the real data from the DRC, we wish to determine which provinces warrant the more complicated seroprevalence model from Sec. II A that includes a constant in addition to the Weibull function, and which are described sufficiently by a constant force of infection alone. To do this, we calculate the Bayes factor between the two models, which, given that our prior belief in the two models is equal, is simply the ratio of the evidence for each of the models. The evidence is the fully marginalized likelihood (FML), and can often be quite difficult to calculate accurately. Luckily for us, the two models in question are nested - we recover the constant-only model when $\alpha = 1.0$ - and so we can use a technique called the Savage-Dickey density ratio.

Two models, $\mathcal{M}_0$ and $\mathcal{M}_1$, where $\mathcal{M}_0$ is the simpler model, are nested when there exists a parameter ω for which $\mathcal{M}_0 = \mathcal{M}_1$ when $\omega = \omega_0$. (For us, constant is the same as Weibull plus constant when the parameter α is equal to 1.0). In this case, the Bayes factor between $\mathcal{M}_0$ and $\mathcal{M}_1$ is simply

$$BF_{01} = \frac{\text{posterior}(\omega = \omega_0)}{\text{prior}(\omega = \omega_0}, \quad (8)$$

where posterior() and prior() refer to the posterior and prior densities at $\omega = \omega_0$. To evaluate this quantity, it is only necessary to generate the posterior distribution of ω, and then compare the posterior and prior densities.

We generate these distributions for each province independently using Markov chain Monte Carlo techniques, described in the next section, and find that only three provinces are adequately described by the constant force of infection model: Bandundu, Kasai Occidental, and Maniema. The rest of the provinces require the more complicated Weibull plus constant model. For all of the results presented in Sec. III, we use these models as noted here for each of the provinces.

Parameter	Prior
α_v, α_f	$\Gamma(2, 500)$
β_v, β_f	$\Gamma(2, 5)$
η_v, η_f	$\Gamma(2, 15)$
γ	$\beta(16, 5)$

TABLE I. Prior distributions for the seven model parameters. $\Gamma(a, b)$ is the gamma distribution with shape a and scale b, and $\beta(c, d)$ is the β distribution with shape parameters c and d.

C. MCMC Sampler

Given our set of model parameters and the definition of the likelihood in Eq. 7, we can generate samples of the posterior distributions of the model parameters using Markov chain Monte Carlo (MCMC) techniques. We use the `PTMCMC` sampler package in Python [1], which incorporates parallel tempering, differential evolution, and proposals along the eigenvectors of the covariance matrix. This sampler is described in detail in [2].

We must specify prior distributions on each of our model parameters before running. These are listed in Table I.

We run for 70000 iterations, keeping every 20th point in order to decrease autocorrelation. At the end of each run, we calculate the number of effective samples via thinning by the autocorrelation length, as

$$N_{eff} = \frac{N}{auto}. \quad (9)$$

Both the effective number of samples and the autocorrelation lengths of all chains are shown in Table **??**.

D. Simulated Data

To confirm that we can accurately recover the force of infection and vaccination hazard using the model we have

chosen, we generate simulated datasets consisting of age-specified measles case, vaccination, and serology data using a previously developed, age-structured MSIRV (Maternally immune, Susceptible, Infected, Recovered, Vaccinated) model [3]. We used demographic parameters from UN estimates for the DRC and increased vaccination linearly from 0 to 50% over the first 30 years of a 50 year simulation. R_0 (the number of secondary cases resulting from the introduction of a single infected individual) was assumed to be constant at 15 over the entirety of the simulation. The force of infection (foi) was not directly chosen to be a Wiebull function, but is generated using the specified R_0 and WAIFW (Who Acquires Infection from Whom) matrix that describes social interactions as estimated via the POLYMOD study [4]. This means that we cannot directly compare the recovered Weibull parameters to injected parameters, but we *can* compare the recovered foi curve to the true values, as well as the recovered vaccination efficacy (γ) to the true value.

While the raw MSIRV model output always entails a value for γ of 1, we can simulate scenarios in which γ is lower by drawing false positive vaccination responses from a binomial distribution with the corresponding probability, and adding these to the vaccination data generated from our simulation. For example, given that our simulation has a maximum vaccination rate of 50%, we can simulate a γ of 0.75 by assuming that, on average, 67% of individuals in a given age class will report that they have been vaccinated.

After we generated a full time-series of case, vaccination, and serology data, we then downsampled the simulation results by randomly drawing the same number of observations as are present in the empirical data from the DRC.

E. Data

Description of the data.

F. Comparing Data and Inference

A standard approach for interpreting the results of Bayesian inference would be to inspect the posterior distributions for the model parameters. But the parameters of our model are not, in themselves, particularly interesting - it is the inferred seroprevalence, case distribution, and vaccination probability that we wish to compare with the data. To do this, we generate the posteriors on $\{\eta_f, \alpha_f, \beta_f, \gamma, \eta_v, \alpha_v, \beta_v\}$ as described, and then draw from these posteriors to generate the seroprevalence, case, and vaccination curves as laid out in Sec. II A. Because the vaccine efficacy, γ, is of interest on its own, we also examine the posterior distributions on γ in detail.

In order to assess the impact of different types of data (case, serology, and vaccination data) on our inferences, we run the MCMC package described in Sec. II C on the vaccination and case data alone, the serology and case data alone, and all three datasets together. The results of all of these analyses are discussed below.

III. RESULTS

A. Simulated Data

Although we have chosen a flexible functional form to represent the force of infection and vaccination hazards, we know that that the true functions present in nature are unlikely to be precisely matched by Weibull distributions. This will necessarily lead to some level of biases in our inferences. To understand these possible biases, we analyze simulated data with known foi, vh, and γ (vaccine efficacy), and examine the results.

Figure 1 shows the injected (dashed lines) and recovered (solid (purple or orange) lines) curves for the force of infection and vaccination hazard from simulated data with vaccine efficacy of $\gamma = 0.9$ (orange) and $\gamma = 0.6$ (indigo). The top panel shows the force of infection, and it is very clear from this panel that a Weibull distribution is not a good fit for the true force of infection, which is bimodal. Because of this, we do not expect to be able to accurately infer the simulation parameters (i.e., γ). We examine the inferred values for this parameter in Figure 1.

This figure shows the posterior distributions and assumed values for γ from simulated datasets with four different values of γ. The top panel is generated by analyzing all three datasets, and the bottom panel using only case and vaccination data. In addition to the assumed values and the posteriors, we also show the prior on γ as a yellow line. In the top panel we see, as expected, that the inferred values for γ are systematically biased away from the injected values - they turn out to be consistently biased low in this case. Although this means that we do not necessarily infer an accurate value for vaccine efficacy with this data using this model, it is worth noting that we do infer the correct *ordering* for γ. That is, if these were four different locations, we would correctly identify the location with the lowest vaccine efficacy.

The bottom panel in this same Figure shows the inferred values for vaccine efficacy using only case and vaccination data - i.e. leaving out the serology data from the analysis. For the higher levels of γ, the results are dominated by the prior. For the lower values, we are again biased quite low, but also again recover the correct ranking of vaccine efficacy.

Finally, we wish to explore the recovered fits to the true data when analyzing only the case and vaccination data. These results are shown in Fig. **??**. Here, we show the recovered (lines) vaccination prevalence, seroprevalence, and case distribution for simulated data with vaccine ef-

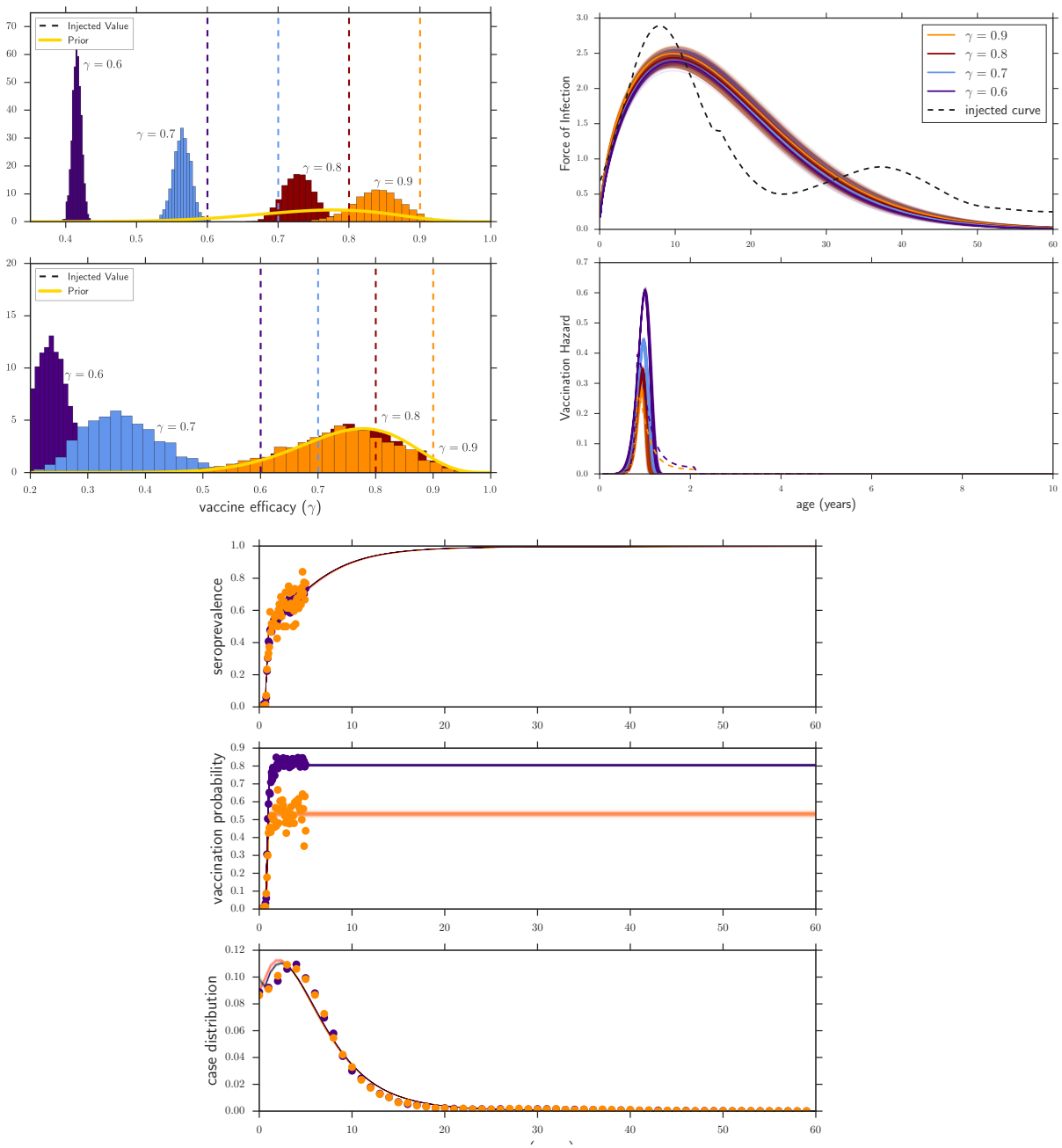

FIG. 1. Results from study using simulated data. The top left plot shows the posterior distributions for each of four different datasets generated with four different vaccine effectiveness (γ). The injected values are shown as dashed lines, and the recovered distributions are labelled with the associated value of γ. The prior is shown as a solid yellow line. The top panel of this plot is generated using all three datasets, and the bottom panel uses only vaccination and serology data. We find that the recovered value of γ is typically biased low, but that the ordering of datasets by γ is recovered correctly. In the cases of high γ and no serology data, the recovered posteriors are dominated by the prior. The top right plot shows the injected (dashed) and recovered (solid) force of infection (top panel) and vaccination hazard (bottom panel) for the same four datasets. The Weibull is not a good fit to the injected foi, but the resulting fits to the observable data (bottom plot) are not overly biased. The fit to the vaccination hazard is similarly not perfect, and the overly-high peaks can explain the general biasing of γ to low values. The bottom plot shows the injected (points) data and recovered (lines) curves for the highest (orange) and lowest (indigo) values of γ for seroprevalence data (top), vaccination data (middle), and case data (bottom).

ficacy of $\gamma = 0.6$ (indigo) and $\gamma = 0.9$ (orange). The true values for each of these are shown as points. We can see that the case and vaccination data are fit well by the recovered model parameters - unsurprising, given that this is the data that was used in the fitting analysis. The seroprevalence data is also fit well in the case of $\gamma = 0.9$, which recall is one of the values for which the posterior on γ is essentially the same as the prior, which includes a lot of weight at the true value. The seroprevalence when $\gamma = 0.6$ is much more biased, although it still follows some of the qualitative features of the true data. Thus for values of vaccine efficacy that are near to the peak of our prior distribution, the seroprevalence inferred using only vaccination and case data is quite accurate. For lower vaccine efficacies, though, seroprevalence data is almost certainly needed to achieve an accurate picture.

B. Application to DRC Data

IV. DISCUSSION

Discussion:

Application for estimating seroprevalence in settings where a sero-survey is not possible ? serosurvey may not be feasible to do frequently, but this allows us to use passively collected age-specific case incidence ? doing a very wide serosurvey may not be feasible, or practical, but case surveillance lets you see signal across a wide age range

What is the use of this? ? estimating seroprevalence in the absence of surveys ? effect size estimate for designing sero-surveys

Discuss the interpretation of the patterns seen in the DRC data ? which places have good vaccination, which have high transmission ? As I recall, some provinces have very wide FOI hazards, indicating that some people are getting infected relatively late in life. There are lots of possible reasons for this, so should give it some thought. Could relatively low exposure, so some people are just not exposed until late in life. Could also reflect inflow of internally displaced people (IDPs). Speculation depends on the pattern.

Several things that we haven?t dealt with here, that we?ll need to at least acknowledge: - we haven?t dealt with the fact that incidence and vaccination are not stationary - we haven?t dealt with supplemental campaigns - these are all things that can be dealt with at least theoretically

Take-home message about merging serosurveillance and case-based surveillance

[1] Justin Ellis and Rutger van Haasteren. jellis18/ptmcmcsampler: Official release, October 2017.

[2] Z. Arzoumanian et al. Gravitational Waves From Individual Supermassive Black Hole Binaries in Circular Orbits: Limits From the North American Nanohertz Observatory for Gravitational Waves. *Astrophys. J.*, 794(2):141, 2014.

[3] C. J. E. Metcalf, J. Lessler, P. Klepac, A. Morice, B. T. Grenfell, and O. N. Bjrnstad. Structured models of infectious disease: Inference with discrete data. *Theoretical Population Biology*, 82(4):275–282, December 2012.

[4] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.*, 5(3):e74, Mar 2008.

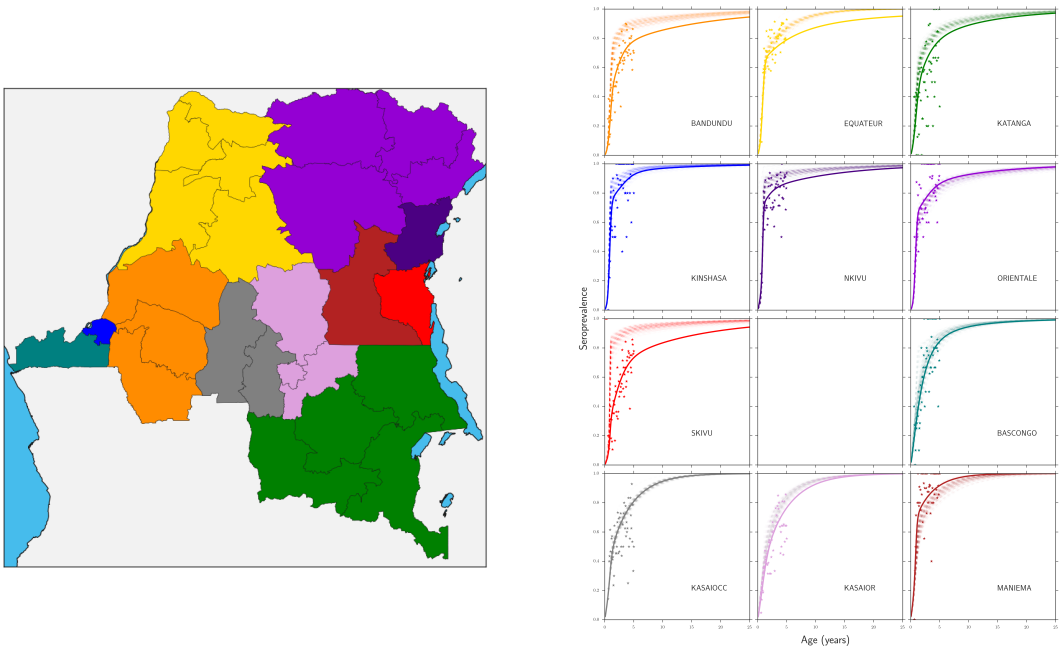

FIG. 2. caption

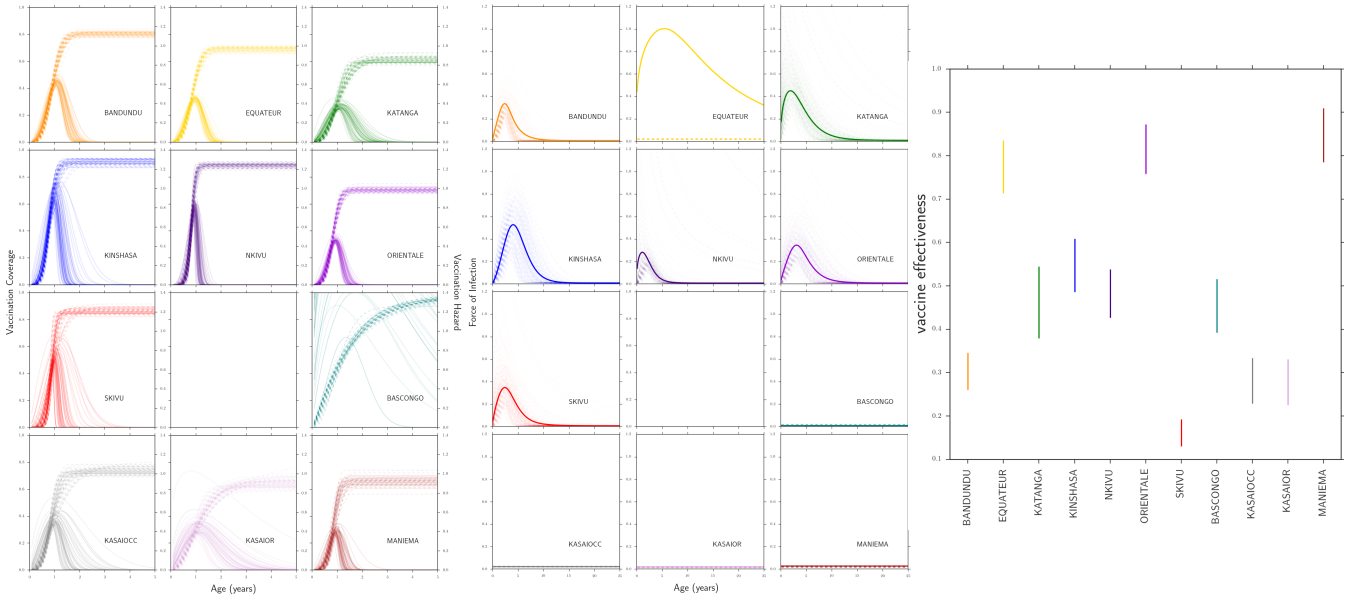

FIG. 3. caption

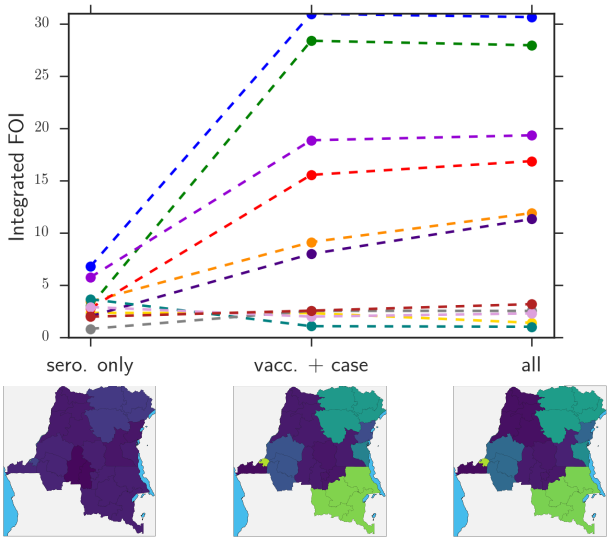

FIG. 4. caption