

Catchy Title Goes Here

Laura Sampson¹ and Matthew Ferrari¹

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 the 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 population 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 processes 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 susceptible 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 and 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](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-specific 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 simplifying 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-specific 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 variation 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 (see 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-prevalence 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

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

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

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

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 [?]. Thus $f(a) \rightarrow f(a|\psi)$ and $v(a) \rightarrow v(a|\theta)$, 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

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

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

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

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

This is the model that we use to analyze simulated data (see Sec. ??), but we found after initial exploration using the data from the Democratic Republic of Congo that the resulting foi was qualitatively quite flat, and so we decided to extend this model by adding a constant to the force of infection as

$$\text{foi}(a) = \omega * \text{Weibull} + (1 - \omega) * \kappa, \quad (5)$$

where ω ranges between 0 and 1, and determines the relative weighting of the Weibull and constant terms, and κ is a constant that is allowed to range between 10^{-4} and 10^2 . The expression in Eq. 4 is then trivially updated to reflect the integral of this more complicated force of infection function. We use Bayesian model selection, discussed in Sec. II B, to determine for which provinces this more complicated model is appropriate.

Eq. 4 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

$$p(\text{vaccinated}|a) = 1 - \exp\left\{-\eta_v (1 - \exp(-(a/\alpha_v)^{\beta_v}))\right\}. \quad (6)$$

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

$$p(\text{case}|a) = \left\{ 1 - \exp \left[-\eta_f \left(\frac{\beta_f}{\alpha_f} \right) \left(\frac{a}{\alpha_f} \right)^{\beta_f - 1} \right] \exp \left[(a/\alpha_f)^{\beta_f} \right] \right\} \times \{1 - p(\text{immune}|a)\}. \quad (7)$$

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

$$\begin{aligned} \log \mathcal{L}(\mathbf{c}, \mathbf{v_t}, \mathbf{v_o}, \mathbf{s_t}, \mathbf{s_o} | \eta_f, \alpha_f, \beta_f, \gamma, \eta_v, \alpha_v, \beta_v) \\ = \sum_a B(v_o(a), v_t(a); p(\text{vaccinated}|a)) \\ + \sum_a B(s_o(a), s_t(a); p(\text{immune}|a)) \\ + M(\mathbf{c}; p(\text{case}|a)), \end{aligned} \quad (8)$$

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

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

and \mathcal{M}_1 is simply

$$BF_{01} = \frac{\text{posterior}(\omega = \omega_0)}{\text{prior}(\omega = \omega_0)}, \quad (9)$$

where $\text{posterior}()$ and $\text{prior}()$ 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: Bas Congo, Kasai Occidental, Kasai Orientale, 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, except for Equateur. In Equateur, the more complicated model was preferred when fitting all three datasets, but the constant only model was preferred when fitting only case and vaccination data, and so these models were used accordingly.

C. MCMC Sampler

Given our set of model parameters and the definition of the likelihood in Eq. 8, 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 150000 iterations, keeping every 100th point after a burnin period of 5000 steps in order to decrease autocorrelation. The two-dimensional posterior distributions for the parameter pairs are shown in Appendix ??.

D. Simulated Data

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

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 .

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 Weibull 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

The real-world data that we apply our model to is from the Democratic Republic of the Congo (DRC), aggregated by province. We note that DRC changed from an 11 province system to a 26 province system in 2015; here we use the 11 province designations. For each province, we have three data sets: vaccination coverage, seroprevalence, and case data, all broken down by age.

The vaccination coverage data is from age-specific survey responses from the 2013-2014 Demographic and Health Survey (DHS) made publicly available by ICF International (The DHS Program - Data Available at <http://dhsprogram.com/Data/>). The survey contains one record for each interviewed woman's child below 5 years of age at the time of the survey. For each child we extracted the age, in months, at the time of the survey and whether or not the child had ever been vaccinated for measles, based on either the record on a vaccination card or parent/guardian recall. We note that parental recall of vaccination may suffer from biases, particularly for older children. Further, the DHS survey question did not distinguish between different sources of measles vaccination (e.g. through routine program, supplemental vaccination campaigns, or outbreak response vaccination campaigns). However, as vaccination cards were rarely available in this survey, parental recall allows the inclusion of many more survey responses; in the 2013-14 DHS in DRC only 1436 respondents had a clearly marked vaccination card, out of 10366 children. Parental recall has previously been shown to provide a relatively robust indicator of vaccination status[5]. Administered measles vaccines have shown to be highly effective in generating protective immunity[6, 7], and we characterize our vaccine effectiveness parameter γ as the probability that a child reported as vaccinated based on either a written record or parental recall was in fact vaccinated and developed protective immunity as a result. Vaccination bias may be lower than one due to recall bias in parental reporting of children's vaccination status, failure of measles vaccination in generating protective immunity, or immunity due to natural infection having already been attained prior to infection.

The seroprevalence data is from dried blood spot samples that were taken from children in a subset of households during the 2013-14 DHS survey and tested for the presence of measles IgG antibodies. Samples were collected from 8267 children between the ages of 0 and 59 months of age. Due to inconclusive results or errors in testing 307 samples were discarded, leaving 7960 samples classified as measles IgG positive or negative. As with the vaccination responses, each sample was geolocated to the cluster-level and we aggregated all samples within each of the 11 provinces.

The age distribution of measles cases was taken from individual case-reporting for DRC reported to the World Health Organization (WHO) between 2011-2017. During this time period, 33098 suspected measles cases were reported, of which 5495 were classified as lab-confirmed by the presence of IgM measles antibodies and 13858 were classified as epidemiologically linked because they occurred close in time and space to a lab confirmed case. For each suspected case we extracted the age, in months, and the province in which the case was recorded. Because the age of many cases was binned to the year (e.g. recorded as 24 or 36 months) we rounded all cases up to the nearest year; thus cases were binned into 1 year age

classes following the convention $(0,1], (1,2]$, etc.

F. Comparing Data and Inference

A standard approach for interpreting the results of Bayesian inference would be inspecting 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 for this parameter 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, vaccination hazard, and γ (vaccine efficacy), and examine the results.

Figure 1 shows the results from this analysis. In the top right panel, we show the injected (dashed lines) and recovered (solid lines) curves for the force of infection and vaccination hazard from simulated data with vaccine efficacy ranging between $\gamma = 0.9$ and $\gamma = 0.6$. The top graph in this panel shows the force of infection, and it is very clear 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 the top left panel of this same figure.

This panel shows the posterior distributions and assumed values for γ from simulated datasets with four different values of γ . The top plot is generated by analyzing all three datasets, and the bottom plot 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 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 important to note 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 plot in this same panel 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, the recovered values 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 all of the simulated datasets. These results are shown in the bottom panel of Fig. 1. Here, we show the recovered (lines) seroprevalence, vaccination probability, and case distribution for simulated data with vaccine efficacy 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 injected data is recovered well with our model, even though the foi and vaccination hazard are not fit particularly well themselves. This is because the data are a function of the *integrated* foi and vaccination hazard, which allows for some ambiguity in the details of the functions themselves.

B. Application to DRC Data

We now turn our attention to the results of our analyses when using the real data from the DRC. As discussed in Section II B, we first determined which model for foi was preferred for each of the eleven provinces we used for our geographical units. We found that when all of the data was included, the only provinces that were better-described by the simple, constant-only model were Bas-Congo, Kasai Occidental, Kasai Orientale, and Maniema. When we looked at the case and vaccination data only, Equateur was also better-described by this model, and when we analyzed only the seroprevalence data, the analysis for all provinces preferred this model.

The left panel of Fig. 2 shows the color-coded provinces. We use these colors throughout this section to show the results from individual provinces. The right panel shows the recovered seroprevalence data (right panel) for all 11 provinces, using all three datasets (solid line) and the vaccination and case data only (dashed lines). The solid line is the mean of the posterior distribution generated using all three datasets, while the dashed lines show the full range of the posterior for the vaccination/case data analysis. We see that, in general, the seroprevalence data leads to a *lower* estimate of seroprevalence in the population than the case and vaccination data alone. This is because the prior on vaccine effectiveness is peaked around 0.85 - that is, when a study participant reports that they or their child is vaccinated, we give an 85% probability

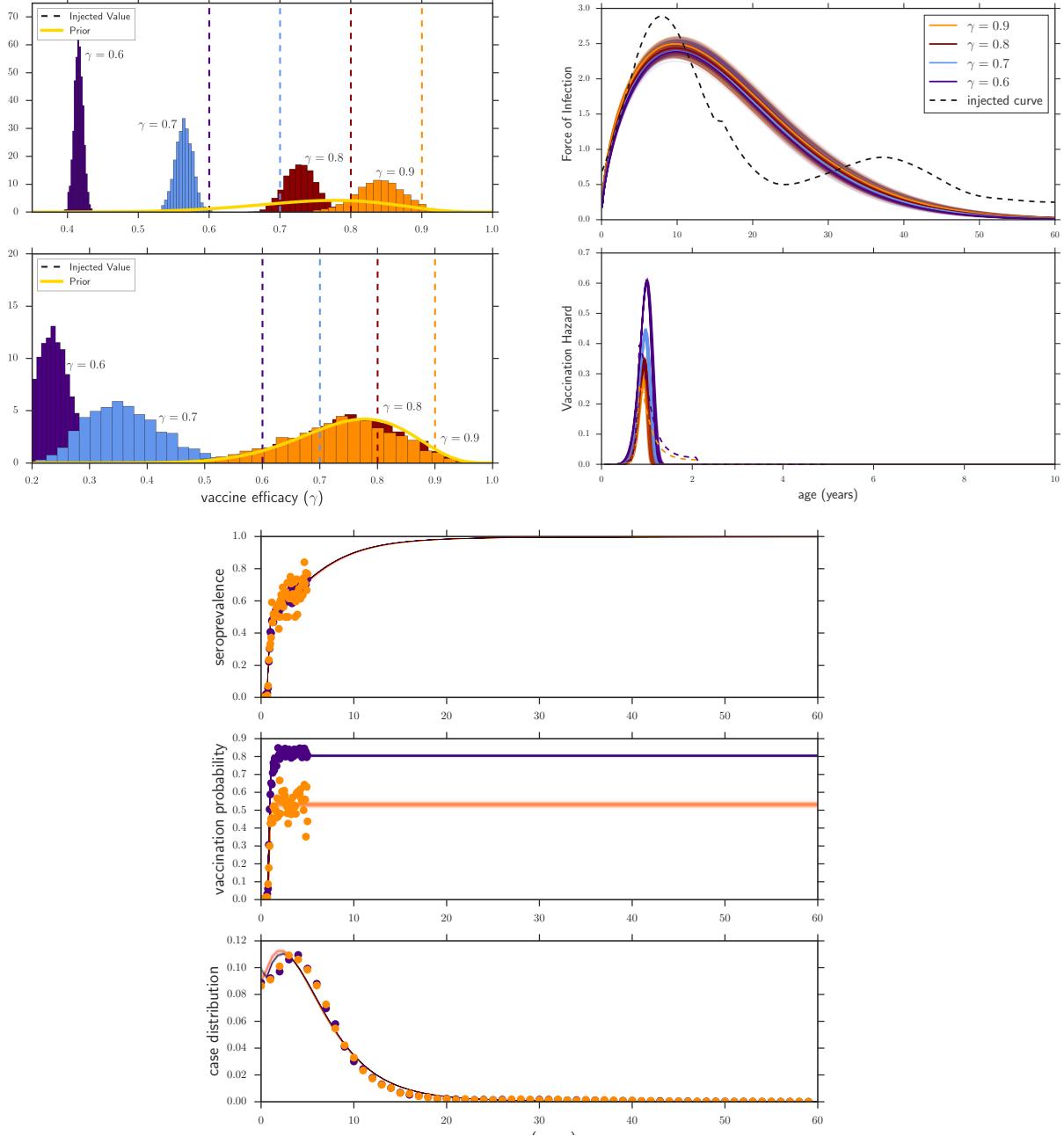


FIG. 1. Results from study using simulated data. The top left plot shows the posterior distributions for each of four different datasets generated with four different vaccine effectiveness (γ). The injected values are shown as dashed lines, and the recovered distributions are labelled with the associated 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).

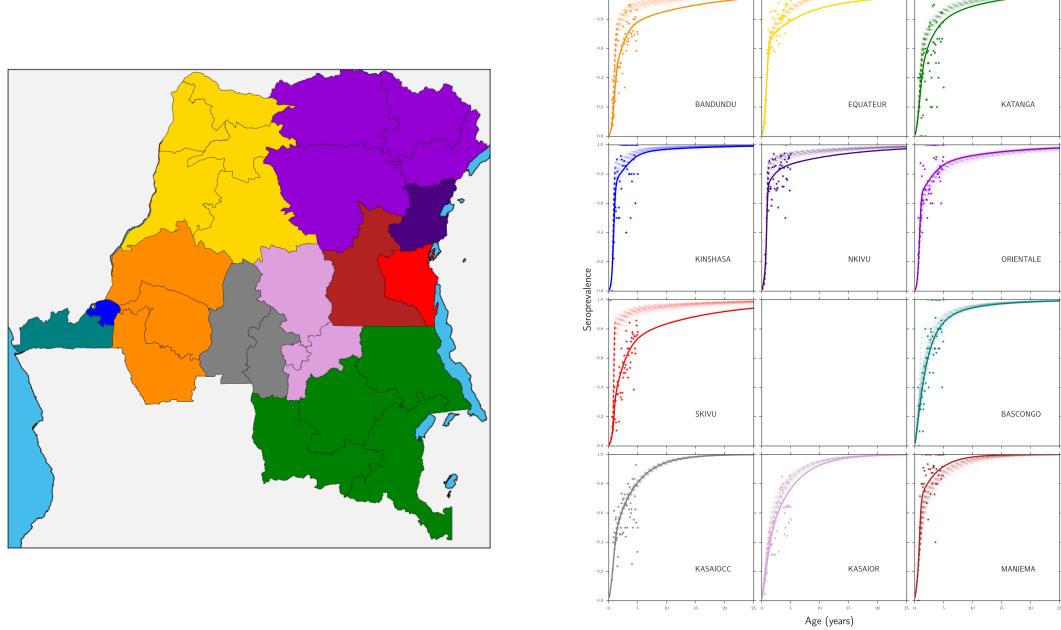


FIG. 2. Left panel: color coding of locations used in our analysis. The data was split into the 11 provinces in existence between 1997 and 2015. The map shows these provinces in different colors, as well as the current administrative boundaries as black lines. The colors on this map correspond to the colors used for each province in all figures in this section. Right panel: seroprevalence data (points) and inferred values up to 25 years. The solid line shows the mean of the posterior distribution for the inferred seroprevalence using all three datasets, while the dashed lines show the full range of the posterior distribution generated using only vaccination and case data. The provinces before the blank space were fit using the Weibull plus constant model for force of infection, while those after (starting with Bas Congo) were fit using only a constant. For the provinces in which seroprevalence data makes a difference in the inferred values, the inferred curve is typically lower than that generated using only vaccination and case data. This is especially pronounced in Sud Kivu.

that not only is the recollection correct, but that the vaccine produced immunity. Without seroprevalence data, there is not enough information in the datasets to pull the value for γ away from this peak, and so we overestimate the immunity in the population. This is particularly noticeable in the results for Sud Kivu. The only province for which this is not the case is Maniema, but looking at the seroprevalence data for Maniema reveals many age groups for which 100% of the population are seropositive. This suggests either a small number of participants, which could be taken into account in future analyses, or something that went wrong with the data reporting.

In the next figure, Fig. 3, we examine the vaccination hazard and probability, the force of infection, and the vaccine efficacy for each of the eleven provinces, inferred using all three datasets. The left panel shows the vaccination data as points, the vaccination hazard as solid lines, and the vaccination probability as dashed lines. All curves are generated from samples drawn from the posterior distributions of the model parameters. For almost all of the provinces, the vaccination hazard is peaked tightly around approximately one year in age, as expected. The height of the vacc. hazard corresponds closely to the final level of vaccination probability - higher vaccination

hazard means a higher probability of getting vaccinated. The only province that does not follow this pattern is Bas Congo, for which there is evidence for a vaccination hazard with non-negligible weight at higher ages. That is, the probability of being vaccinated is still increasing at the right edge of this graph, which corresponds to five years old.

The next (middle) panel in Fig. 3 shows the force of infection recovered for each province using all data (solid) and vaccination/case data (dashed). For all provinces except Equateur, the force of infection is either Weibull plus constant (first seven cells), or constant-only (final four cells) for both datasets. In Equateur, the simpler model is used when vaccination/case data only are used. We see here that all provinces have a non-zero force of infection at higher ages.

The final (right) panel in Fig. 3 shows the mean value and uncertainty for vaccine efficacy ($\bar{\gamma} \pm \sigma$) for all eleven provinces, with the same color-coding as has been used throughout this section. As expected from the serology results, Sud Kivu has the lowest vaccine efficacy. Because of our results using the simulated data, we believe that the ordering of provinces by γ is correct - but we do not believe that the actual values should be taken seriously. From that analysis, we suspect that the recovered values

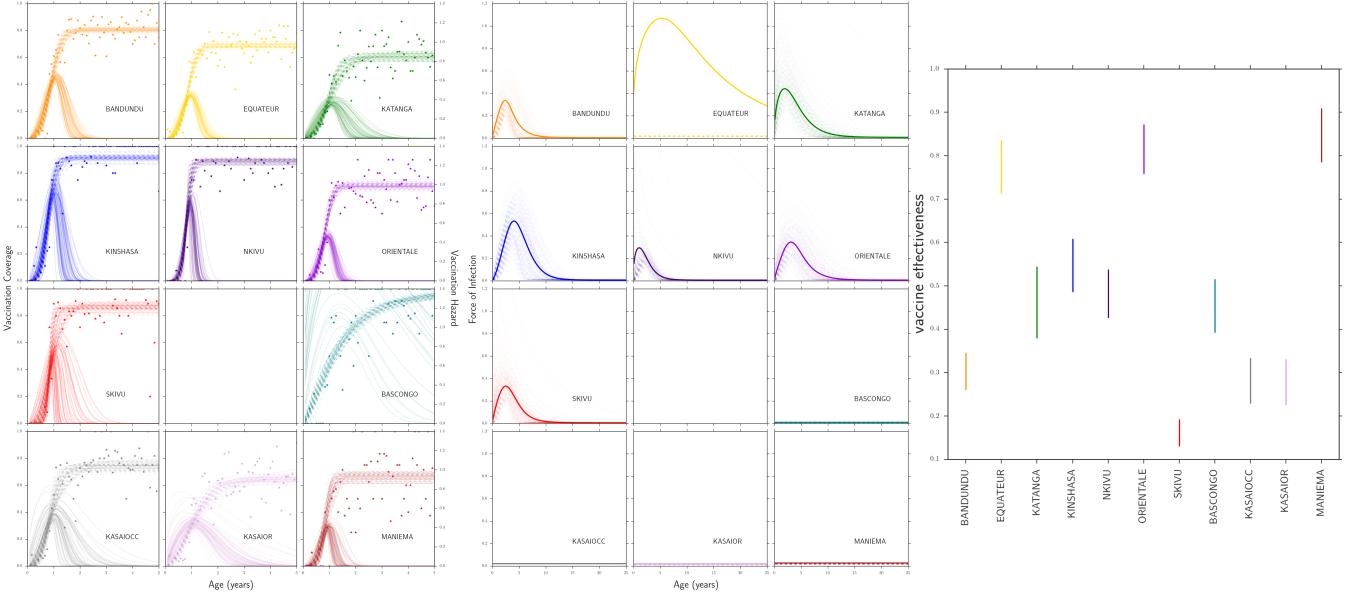


FIG. 3. Left panel: Vaccination hazard (solid lines, right y-axis) and vaccination coverage (dashed lines, left y-axis) recovered using all three sets of data from all 11 provinces, generated from a set of posterior samples. For most provinces, the vaccination hazard is tightly peaked near one year, as expected. Middle panel: The recovered force of infection using all data (solid curve) and only vaccination and case data (dashed curves). As in Figure 2, the solid curve is the mean of the posterior distribution, and the provinces before the blank space were fit using the Weibull plus constant model while those after were fit using the constant-only model for force of infection. In most cases, the inclusion of the serology data did not make a large difference in the inferred force of infection. Right panel: recovered value for vaccine efficacy ($\pm 1\sigma$) from analyses using all three data sets for each province. Because of our results from the simulated data, we do not expect that these values to reflect the true vaccine efficacy (in general, they are probably biased low), but we *do* expect that the ranking of provinces by vaccine efficacy is correct.

are likely lower than the true values.

Finally, in Fig. 4, we assess what we call the ‘risk’ inferred for with each province, which we define as

$$r = \int_0^{10} \text{foi}(a)da, \quad (10)$$

or the force of infection integrated from 0 to 10 years. We calculate the foi using serology data only (left-most point), case and vaccination data only (middle point), and all three datasets (right-most point), and show the integrated values in Fig. 4. The provinces are color-coded as throughout this section, and a dashed line is drawn to guide the eye between the results for different datasets. Below the points, we show maps of the DRC with each province colored according to the risk level, with dark blue being lowest risk and yellow highest. The most obvious conclusion from this plot is that the inclusion of case and vaccination data leads us to drastically increase our assessment of risk in many provinces. Interestingly, the inclusion of serology data does not make a significant difference in the calculated risk. The information from serology data seems to be most useful for inferring vaccine efficacy, rather than force of infection.

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

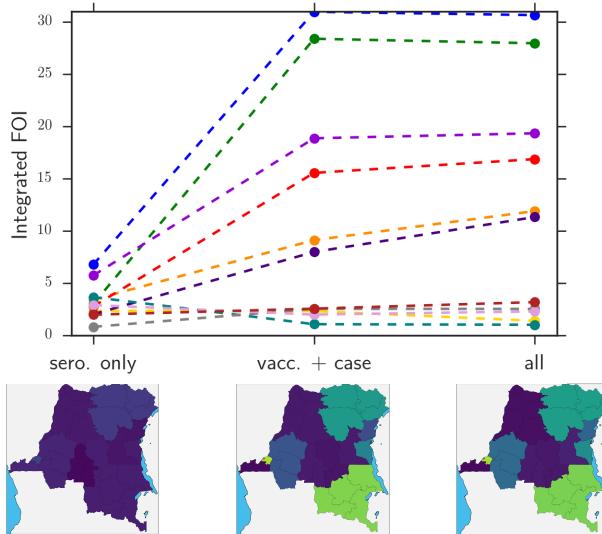


FIG. 4. Force of infection integrated to age ten ('risk') for all eleven provinces, color coded as indicated in Figure 2, calculated using the inferred foi from only serology data (left-most point), case and vaccination data (middle point), and all three datasets. Both the ranking of provinces by risk and the overall values change dramatically between the serology-only and the other two points, but the difference between vaccination plus case and all three data sets is minimal. The maps show the risk in all provinces, with yellow being high risk and dark blue being low risk.

- these are all things that can be dealt with at least theoretically

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

*

Appendix A: Full Results from MCMC Runs

Here, we show the two-dimensional posterior distributions for pairs of parameters from the MCMC runs on the DRC data.

-
- [1] Justin Ellis and Rutger van Haasteren. *jellis18/ptmcmsampler*: 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. Bjørnstad. 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.
 - [5] James Ndirangu, Ruth Bland, Till Brnighausen, and Marie-Louise Newell. Validating child vaccination status in a demographic surveillance system using data from a clinical cohort study: evidence from rural South Africa. *BMC Public Health*, 11:372, May 2011.
 - [6] Dalya Gris, Jill Mccready, John C. Watson, William L. Atkinson, Janet L. Heath, William J. Bellini, and Anthony Pollio. Measles Vaccine Effectiveness and Duration of Vaccine-induced Immunity in the Absence of Boosting from Exposure to Measles Virus. *The Pediatric Infectious Disease Journal*, 15(12):1082, December 1996.
 - [7] G. E. King, L. E. Markowitz, P. A. Patriarca, and L. G. Dales. Clinical efficacy of measles vaccine during the 1990 measles epidemic. *The Pediatric infectious disease journal*, 10(12):883–888, December 1991.