Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Ecological inference for infectious disease data, with application to vaccination strategies

Leigh H. Fisher*[1] | Jon Wakefield[2,3]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Washington, USA
[2]Department of Biostatistics, University of Washington, Washington, USA
[3]Department of Statistics, University of Washington, Washington, USA

Correspondence
*Email: lfisher@fredhutch.org

Abstract

Disease surveillance systems provide a rich source of data regarding infectious diseases, aggregated across geographical regions. The analysis of such ecological data is fraught with difficulties, and unless care, and suitable data summaries are available, will lead to biased estimates of individual-level parameters. We consider using surveillance data to study the impacts of vaccination. To catalog the problems of ecological inference we start with an individual-level model, that contains familiar parameters, and derive an ecologically consistent model for infectious diseases in partially vaccinated populations. We compare with other popular model classes and highlight deficiencies. We explore the properties of the new model through simulation, and demonstrate that under standard assumptions, the ecological model provides less biased estimates. We then fit the new model to data collected on measles outbreaks in Germany from 2005–2007.

KEYWORDS:
Count data; Ecological bias; Time series; Vaccine coverage

1 | INTRODUCTION

A wide range of diseases are monitored at the local, state, and national levels using disease surveillance systems designed to assess the current disease burden or to detect emerging outbreaks. Although there are a variety of approaches to surveillance, ranging from daily collection of de-identified electronic medical records to mandatory reporting of certain notifiable diseases, the resulting data typically captures information for large populations over time. For this reason, disease surveillance systems are frequently a primary source of information for public health researchers and officials who use such data to design and deploy effective interventions. While this approach is economical, cases are typically aggregated in space, time, or both and the information regarding any single case is limited.

This aggregation can present challenges when studying the spread of infectious disease. In the social sciences and non-infectious disease epidemiology, aggregated data is often analyzed with established disease mapping approaches such as ecological regression. However, the risk of drawing erroneous individual-level conclusions from group-level data has been well characterized.[1–6] This phenomenon is referred to as ecological bias and can arise when the form of the risk model changes under aggregation. When the model for the individual-level risk of disease is a nonlinear function of the exposure, as is typically the case for infectious disease models, the form of the marginal aggregate risk model changes as a result of the within-group variability of the exposure that is not accounted for in the group-level model.[7]

In the infectious disease setting, ecological regression approaches are typically not considered since they do not leverage known dependencies. For aggregated infectious disease data, there are two common approaches in the literature: the time series SIR (TSIR) model[8] and the epidemic-endemic framework.[9–15] Under the TSIR approach, the number of susceptible and infected individuals are modeled independently without recourse to a development from the individual-level. The epidemic-endemic framework is motivated by spatial branching processes, and is closely related to standard SIR and multivariate time series SIR models.[16] While these epidemic-endemic models are easily fit in standard software via the `surveillance` package in the `R` programming environment.[14] A recent review compares and contrasts the two classes of models.[17] For both approaches to modeling aggregated infectious disease data, the risks of infections are nonlinear and thus inference is susceptible to ecological bias. However, there has been little discussion of ecological bias for aggregate infectious disease models.[18] In particular, the ecological aspects of the epidemic-endemic model have not been investigated.

In this manuscript, we consider using aggregated surveillance data to study the impact of vaccination on infectious disease transmission. To avoid ecological bias, we start with an individual-level infectious disease model that includes vaccination and derive an ecologically consistent infectious disease model for a partially vaccinated populations. This ecological vaccine model is easily fit, and provides estimates of familiar epidemiological parameters. The remainder of this paper is organized as follows: in Section 2 we motivate the aim of the paper and introduce some notation and preliminary concepts. In Section 3 we develop an ecologically consistent vaccine model under two models of vaccine action; we present simulations to better understand the behavior of the ecological vaccine model in Section 4; and fit the ecological vaccine model to measles data from Germany in Section 5. Final comments appear in Section 6.

2 | MOTIVATION AND NOTATION

In surveillance data, new cases are commonly reported in discrete time and space. It is common to use time steps relative to the disease of interest, meaning that we are assuming the sum of incubation and infectious times is approximately that of the observation times. For example, for measles, the data are often aggregated over 2 week periods. We denote the number of cases

and the population size for area i and time t by Y_{it} and N_{it}. Let S_{it} denote the number of susceptible individuals and x_{it} the proportion of vaccinated individuals in area i and time t. Area- and time-specific covariates other than vaccination coverage are denoted z_{it}.

Recently, the epidemic-endemic model was derived via aggregation of an individual-level model, and the framework was extended to handle a stratified population.[15] We briefly review this derivation before discussing how the epidemic-endemic models are typically applied when considering vaccination. We use $\lambda_{it}^{\dagger}$ to denote the generic force of infection, or the risk of an individual who was susceptible at time $t-1$ becoming infected by time t in area i.[19] Assuming a constant hazard of infection between time steps, the probability of a susceptible individual in area i and time $t-1$ becoming infected by time t is determined by the hazard rate $\lambda_{it}^{\dagger}$, implying the following individual-level model: Pr (infection in $(t-1,t]$ | no infection by $t-1$, area i) $= 1-e^{-\lambda_{it}^{\dagger}}$. A Reed-Frost chain binomial SIR model is implied if we additionally assume that the time until infection is independent for all susceptible individuals[19]; hence, the number of new infectives in area i at time t can be modeled as $Y_{it}|\lambda_{it}^{\dagger} \sim \text{Binomial}\left(S_{i,t-1}, 1-e^{-\lambda_{it}^{\dagger}}\right)$. When $\lambda_{it}^{\dagger}$ is small, the Taylor expansion, $1-\exp(-\lambda_{it}^{\dagger}) \approx \lambda_{it}^{\dagger}$, simplifies the form of the probability of infection. When the number of susceptibles, $S_{i,t-1}$ is large, and the probability of infection is small, the binomial distribution can be approximated by a Poisson distribution so that $Y_{it}|\mu_{it} \sim \text{Poisson}\left(\mu_{it}\right)$, where $\mu_{it} = S_{i,t-1}\lambda_{it}^{\dagger}$. When the number of new infections is small and the population is large, the number of susceptibles can be approximated by the initial number of susceptibles, $S_{it} \approx N_i$.

In the infectious disease setting, there are typically multiple sources of infection. For example, a susceptible may become infected from an infective in their own area, another area, or from an an environmental reservoir or infective external to the study region. Typically, the epidemic-endemic framework decomposes the force of infection into three components: autoregressive (AR), neighborhood (NE), and endemic (EN), where the endemic component includes all other sources of infection.[9] For simplicity here, we consider the AR and EN components only (though the discussion holds if neighborhood terms are included also). Considering a competing risk framework, we can write $\lambda_{it}^{\dagger} = \lambda_{it}^{\text{AR}\dagger} + \lambda_{it}^{\text{EN}\dagger}$, where $\lambda_{it}^{\text{AR}\dagger}$ and $\lambda_{it}^{\text{EN}\dagger}$ are generic forms of the component-specific risks. A frequency dependent transmission model implies $\lambda_{it}^{\text{AR}\dagger} = \lambda_{it}^{\text{AR}} y_{i,t-1}/N_i$.[15] Then assuming rare events and $S_{it} \approx N_i$, we obtain a general form of the epidemic-endemic model, with $Y_{it}|\mu_{it} \sim \text{Poisson}\left(\mu_{it}\right)$, where

$$\mu_{it} \quad = \quad S_{i,t-1}\lambda_{it}^{\dagger} \quad = \quad \underbrace{\lambda_{it}^{\text{AR}}\, y_{i,t-1}}_{\text{Autoregressive}} \quad + \quad \underbrace{\lambda_{it}^{\text{EN}} N_i}_{\text{Endemic}} . \tag{1}$$

The autoregressive component accounts for the disease risk from infectives in the previous time period and in the same area. The endemic component describes the additional risk from environmental reservoirs that contribute to the risk of infection or other sources of infection not already accounted for by the other component(s). The parameters λ_{it}^{AR} and λ_{it}^{EN} are rates and determine the relative contributions of cases from the respective sources, though are not directly comparable.

The epidemic-endemic framework typically models the number of cases in area i and time t with a negative binomial distribution with mean $\lambda_{it}^{\dagger}$.[14] For simplicity, we model overdispersion via a Poisson distribution with log-normal random effects, with the mean decomposed as in equation (1). Each component can be modeled with a log-linear model to include covariates as well as fixed and random effects. For example, the autoregressive component may take the form

$$\log \lambda_{it}^{\mathrm{AR}} = \alpha_{\mathrm{AR}} + a_i + \boldsymbol{\beta}^{\mathrm{AR}} \boldsymbol{z}_{it}, \tag{2}$$

where α_{AR} is a log-risk intercept, a_i are area-specific fixed (or random) effects, $\boldsymbol{z}_{it}$ are area- and time-specific covariates, and $\exp(\boldsymbol{\beta}^{\mathrm{AR}})$ are the associated covariate relative risks. The endemic component can be modeled in a similar fashion to the above auto-regressive component. Seasonality can be included in either component of the model by adding to the log-risk, a term of the form, $\sum_{s=1}^{S} \left[\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)\right]$, where S is the number of pairs of sines and cosines to include and ω_s are Fourier frequencies. For biweekly data, $\omega_s = 2\pi s/26$. In practice, seasonal terms have been included in only the endemic component.[12–14]

In the `surveillance` package, parameter estimates are quickly obtained for the epidemic-endemic models via penalized maximum likelihood estimation.[14] While epidemic-endemic models can be used for prediction, they are more often used to smooth observed counts, in which case parameter interpretation is not done extensively. Within the epidemic-endemic framework, there has been no discussion of the ecological bias implications of the use of loglinear models of the form (2). Appendix A shows the inconsistency between the individual and ecological models in some simple situations, using this loglinear model, and simulations in Section 4.2 provide numerical examples of ecological bias in this setting.

In the context of studying vaccination on disease spread, there are two analyses that use the epidemic-endemic framework to model measles in Germany that include vaccination coverage.[12,14] Both analyses consider multiple ways of incorporating vaccination coverage into the mean model and use AIC to select a final model. However, the analyses of separate data sets produced different models for measles in Germany. One analysis included vaccination coverage in only the autoregressive component, while the other included it in only the endemic component. For example, the model $Y_{it}|\mu_{it} \sim \mathrm{Poisson}(\mu_{it})$, included vaccination coverage in the endemic component, with

$$\mu_{it} = \lambda_{it}^{\mathrm{AR}}\, y_{i,t-1} + (1 - x_i)^{\alpha_1}\, \lambda_{it}^{\mathrm{EN}} N_{it}/N, \tag{3}$$

where x_i is the proportion of vaccinated individuals in area i.

While this approach may lead to models that fit the data well, it fails to account for the scientific context of how vaccination affects susceptibility. In (3), if the proportion of unvaccinated individuals, $(1 - x_i)$ is thought of as a proxy for the number of susceptible individuals in the population, then the parameter associated with the vaccination coverage, α_1 can be thought of as a flexibility parameter to improve model fit.[12] Moreover, the interpretation of the parameter associated with vaccination coverage

can be cumbersome or non-intuitive and the parameters may not be comparable across analyses. For example, the interpretation from (3) is that the expected multiplicative change in *endemic* incidence associated with a doubling of the proportion of susceptible individuals in area i is estimated to be $2^{\hat{\alpha}_1}$.[14]

Alternatively, it is common to account for vaccination in applications of the TSIR framework by augmenting the susceptibles model to account for vaccination coverage. For example, in the context of modeling hand, foot and mouth disease in China, the basic accounting equation for susceptibles the number of new births is reduced by the vaccination coverage, which is assumed known and vaccine effect is not estimated.[20] As these examples demonstrate, current approaches to modeling aggregate data may be inappropriate when the goal is to study the covariate effects on disease spread. When the interest is to study the effects of vaccination for an imperfect vaccine, the resulting models lack familiar parameter interpretation, and primarily focus on model fit.

Before proceeding, we take a moment to introduce a key parameter for quantifying infectious diseases that is regularly used in practice. The basic reproductive number, represented by R_0, is defined as the average number of individuals a typical infectious individual would infect in a completely susceptible population.[21] When a portion of the population is immune, either because of vaccination or previous infection, the average number of new infections caused by a single infectious is called the *effective reproductive number*, represented by R. In our setting, where x is the proportion of the population that is immune to infection, either through natural infection or vaccination, $R = (1 - x)R_0$. For both R_0 and R, values less than 1 imply that major outbreaks can be avoided.

3 | ECOLOGICAL VACCINE MODEL DEVELOPMENT

3.1 | Introduction

With inference as a primary goal, we now develop an aggregate infectious disease model with vaccination for inference. For clarity, we develop the ecological model in a single area and with a generic force of infection, although as we show in Section 5, extensions to multiple areas and more complex forms of risk can be made. We further assume that the vaccine only affects an individual's susceptibility to infection (and not infectiousness or disease progression) and that vaccination provides lifetime immunity. We let ϕ be the reduction in a vaccine recipient's risk of infection (the vaccine effect) after vaccination, and assume a constant vaccine coverage denoted by x. We subscript the number of susceptibles, cases, and force of infections with v and u to indicate vaccinated and unvaccinated. Hence Y_{ut} and Y_{vt} is the total number of unvaccinated and vaccinated infectives at time t, such that $Y_t = Y_{ut} + Y_{vt}$. We assume that vaccination is administered in a totally susceptible population, so that at $t = 0$, the number of unvaccinated susceptible individuals is $S_{u0} = (1 - x)N$.

To properly model the effects of vaccination, at the population level, it is important to consider how the vaccine reduces an individual's risk of infection. We consider aggregate models for two modes of vaccine action: leaky and all-or-none. Leaky vaccines are assumed to reduce the risk of infection by a constant proportion for all vaccinated individuals; in contrast, all-or-none vaccines provide full protection from infection to vaccinated individuals when successful, but fail to provide protection with some probability.[19] In other words, leaky vaccines reduce the per-exposure risk of infection, while an all-or-none vaccine's protection is independent of the number of contacts made. In reality, a given vaccine may not fall squarely into one of these two categories, but we use these two different models to explore these extremes. In the subsequent sections, we show that regardless of the assumed mode of vaccine action, there is a common, ecologically consistent model that can be fit to aggregate data.

3.2 | All-or-none vaccine ecological model

For an all-or-none vaccine, it is assumed that the vaccine fails with probability $(1-\phi)$ and offers no partial protection in this case.[19] This implies that the number of susceptible individuals who were vaccinated is $S_{v0}(\phi) = (1-\phi)xN$, and $\lambda_{vt} = \lambda_{ut} = \lambda_t^\dagger$ is the common risk of infection. We denote the number of susceptibles at time t by $S_t(\phi)$ to emphasize that the number of susceptibles is a function of the vaccine effect. At time $t = 0$, the number of susceptibles at time $S_0(\phi) = S_{u0}(\phi) + S_{v0}(\phi) = (1-x)N + (1-\phi)xN = (1-\phi x)N$. The number of new infections at time $t+1$ can be modeled as

$$Y_{t+1} \mid \lambda_t^\dagger \sim \text{Binomial}\left(S_t(\phi),\, 1-\exp(-\lambda_t^\dagger)\right), \tag{4}$$

where $S_t(\phi) = S_{t-1}(\phi) - Y_t$. In the rare disease setting, the binomial can be approximated by a Poisson and when $\lambda_t^\dagger$ is small, a Taylor expansion approximates $1-\exp(-\lambda_t^\dagger) \approx \lambda_t^\dagger$ so that $Y_{t+1} \mid \lambda_t^\dagger \sim \text{Poisson}\left(S_t(\phi)\,\lambda_t^\dagger\right)$, where $S_t(\phi) = (1-\phi x)N - \sum_{k=1}^{t} Y_k$. When the susceptible population is sufficiently large, and the number of cases is small, the number of susceptibles is effectively constant and can be approximated by $S_t(\phi) \approx (1-\phi x)N$. The ecological model in (4) becomes

$$Y_{t+1} \mid \lambda_t^\dagger \sim \text{Poisson}\left(\lambda_t^\dagger(1-\phi x)N\right), \tag{5}$$

when the approximations are valid. In Section 4.1, we consider the conditions under which these modeling assumptions are reasonable.

3.3 | Leaky vaccine ecological model

Under the leaky vaccine model vaccinated individuals are still susceptible to infection, and therefore $S_{u0} = (1-x)N$ and $S_{v0} = xN$. Additionally, the leaky vaccine implies that we can write the risk of infection for the vaccinated as a function of that

in the unvaccinated population and the vaccine effect:

$$\lambda_{vt}^{\dagger} = (1-\phi)\lambda_{ut}^{\dagger}. \tag{6}$$

Then, the number of new infections at time $t+1$ can be modeled as

$$Y_{u,t+1} \mid \lambda_{ut}^{\dagger} \sim \text{Binomial}\Big(S_{ut}, 1-\exp(-\lambda_{ut}^{\dagger})\Big), \tag{7}$$

$$Y_{v,t+1} \mid \lambda_{vt}^{\dagger} \sim \text{Binomial}\Big(S_{vt}, 1-\exp(-\lambda_{vt}^{\dagger})\Big), \tag{8}$$

where $\lambda_{ut}^{\dagger}$ is the risk of infection for an unvaccinated susceptible at time t, and $\lambda_{vt}^{\dagger}$ is defined in (6); the number of susceptibles at time $t+1$ are

$$S_{u,t+1} = S_{u,t} - Y_{u,t+1} \qquad \text{and} \qquad S_{v,t+1} = S_{v,t} - Y_{v,t+1}.$$

The resulting aggregate model is a convolution of binomials, where

$$\Pr\left(Y_t = y \,\middle|\, \lambda_{ut}^{\dagger}, \lambda_{vt}^{\dagger}\right) = \sum_{z=0}^{y} \Pr\left(Y_{ut} = z \,\middle|\, \lambda_{ut}^{\dagger}\right) \Pr\left(Y_{vt} = y - z \,\middle|\, \lambda_{vt}^{\dagger}\right). \tag{9}$$

When the susceptible populations or disease counts are large, this aggregate model will be computationally expensive and practically intractable. When the risks of infection are small, the Taylor approximation simplifies the probability of infection in equations (7) and (8). Moreover, when infections are rare, the binomial distributions can be approximated by Poissons. Hence, when risk of infection is small for both the unvaccinated and vaccinated populations, the number of new infections in each group is approximately

$$Y_{u,t+1} | \lambda_{ut}^{\dagger} \sim \text{Poisson}\Big(S_{ut}\, \lambda_{ut}^{\dagger}\Big), \tag{10}$$

$$Y_{v,t+1} | \lambda_{ut}^{\dagger}, \phi \sim \text{Poisson}\Big(S_{vt}\,(1-\phi)\lambda_{ut}^{\dagger}\Big). \tag{11}$$

The resulting aggregate model, when the risk is small for both vaccinated and unvaccinated groups is

$$Y_{t+1} \,\Big|\, \lambda_{ut}^{\dagger}, \phi \sim \text{Poisson}\left(\left(S_{ut} + S_{vt}(1-\phi)\right)\lambda_{ut}^{\dagger}\right). \tag{12}$$

Compared to the convolution model of (9), this likelihood is more tractable in large populations with few cases. However, this model still requires knowing the number of susceptibles by vaccination status, which is typically not known or easily approximated. If it is reasonable to assume that the number of infectives is negligible when compared to the size of the susceptible pool, i.e., $S_{ut} \approx S_{u0}$ and $S_{vt} \approx S_{v0}$, the ecological model for a partially vaccinated population is approximately

$$Y_{t+1} \,\Big|\, \lambda_{ut}^{\dagger}, \phi \sim \text{Poisson}\left(\lambda_{ut}^{\dagger}(1-\phi x)N\right), \tag{13}$$

which is identical to the ecological model derived assuming an all-or-none vaccine given in equation (5).

3.4 | Comments on the ecological vaccine model

We summarize the development of the ecological vaccine model starting from the all-or-none and leaky vaccine assumptions, as well as the simplifying assumptions that result in the ecological vaccine model in Table 1. Both the all-or-none and leaky

	All-or-none	**Leaky**
Initial susceptible population		
	$S_{u0}(\phi) = (1-x)N$ $S_{v0}(\phi) = (1-\phi)xN$	$S_{u0} = (1-x)N$ $S_{v0} = xN$
Force of infection		
	$\lambda^\dagger_{ut} = \lambda^\dagger_t$ $\lambda^\dagger_{vt} = \lambda^\dagger_t$	$\lambda^\dagger_{ut} = \lambda^\dagger_t$ $\lambda^\dagger_{vt} = (1-\phi)\lambda^\dagger_t$
Progression		
$Y_{u,t+1} \mid \lambda^\dagger_{ut}$	$\text{Bin}\left(S_{ut}(\phi),\, 1-e^{-\lambda^\dagger_t}\right)$	$\text{Bin}\left(S_{ut},\, 1-e^{-\lambda^\dagger_t}\right)$
$Y_{v,t+1} \mid \lambda^\dagger_{vt}$	$\text{Bin}\left(S_{vt}(\phi),\, 1-e^{-\lambda^\dagger_t}\right)$	$\text{Bin}\left(S_{vt},\, 1-e^{-(1-\phi)\lambda^\dagger_t}\right)$
Implied aggregate model		
$Y_{t+1} \mid \lambda^\dagger_t$	$\text{Bin}\left(S_t(\phi), 1-e^{-\lambda^\dagger_t}\right)$	Convolution of binomials
Simplifying assumptions		
Poissons approximate binomials		
	$\text{Poi}\left(S_t(\phi)\left(1-e^{-\lambda^\dagger_t}\right)\right)$	$\text{Poi}\left(S_{ut}\left(1-e^{-\lambda^\dagger_t}\right) + S_{vt}\left(1-e^{-(1-\phi)\lambda^\dagger_t}\right)\right)$
Taylor approximation		
	$\text{Poi}\left(S_t(\phi)\lambda^\dagger_t\right)$	$\text{Poi}\left(\left(S_{ut} + (1-\phi)S_{vt}\right)\lambda^\dagger_t\right)$
Negligible number of infections		
	$S_t(\phi) \approx (1-\phi x)N$	$S_{ut} \approx (1-x)N, \quad S_{vt} \approx xN$
Ecological vaccine model		
	$Y_{t+1} \mid \lambda^\dagger_t, \phi \sim \text{Poisson}\left(\lambda^\dagger_t(1-\phi x)N\right)$	

TABLE 1 Summary of the all-or-none and leaky vaccine models and the assumptions for the ecological vaccine model. N is the total population; x denotes the proportion of the population vaccinated (assumed constant over time); ϕ is the vaccine effect on susceptibility; S_{ut} and S_{vt} denote the number of unvaccinated and vaccinated susceptibles at time t; Y_{ut} and Y_{vt} denote new cases in time t among unvaccinated and vaccinated; and $\lambda^\dagger_t$ is a generic force of infection.

vaccine models can be approximated by the ecological vaccine model when the following simplifying assumptions can be made:

1. Poisson approximation to the binomial distribution
2. Force of infection approximation: $1 - e^{-\lambda^\dagger_t} \approx \lambda^\dagger_t$
3. Negligible number of infections: $S_{ut} \approx S_{u0}$ for unvaccinated individuals, and $S_{vt} \approx S_{v0}$ for vaccinated individuals. Note that the number of susceptibles may also be a function of the vaccine effect.

This list of assumptions helps illuminate when the ecological vaccine model we have developed is appropriate to use. The fact that, when aggregated, both vaccine models can be approximated by the same model suggests that with aggregated data, there is not sufficient information to tease apart the mechanism of vaccine protection. In fact, in Appendix B, we derive the ecological vaccine model assuming the vaccine that has both leaky and all-or-none effects and show that the specific vaccine effects are not identifiable with the ecological vaccine model.

4 | SIMULATIONS

4.1 | Assessing the simplifying assumptions in the absence of vaccination

We first assess the conditions under which these simplifying assumptions are appropriate in the absence of vaccination via simulation. Each simulated epidemic starts with a single infected individual in an otherwise susceptible population of $N = 100,000$; in other words let $Y_0 = 1$ and $S_0 = N - Y_0$. And the number of cases over the course of a given epidemic are simulated as follows:

$$Y_{t+1}|\lambda_t^\dagger, y_t \sim \text{Binomial}\left(S_t, 1 - e^{-\lambda_t^\dagger}\right),$$
$$\lambda_t^\dagger = e^{\alpha_{\text{AR}}}\, y_t/N + e^{\alpha_{\text{EN}}},$$
$$S_t = N - \sum_{k=0}^{t} y_k.$$

We simulate epidemics for high, medium, and low values of R_0, which correspond to $\alpha_{\text{AR}} = \log(2.5), \log(1)$, or $\log(0.85)$, and fix $\alpha_{\text{EN}} = -10$. To increase variability in the initial number of cases in each simulated epidemic, we discard observations from $t = 0, \ldots, 4$ and simulate the equivalent of 3 years of weekly data starting from $t = 5$. We simulate 250 epidemics for each of the three simulation scenarios. For each simulated epidemic, we fit models from all possible combinations of the three simplifying assumptions summarized in Section 3.4 (and Table 1) and compare the maximum likelihood estimates (MLEs) obtained via numerical optimization. Specifically, we fit:

1. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(S_t, 1 - e^{-\lambda_t^\dagger}\right)$.
2. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(N, 1 - e^{-\lambda_t^\dagger}\right)$.
3. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(S_t, \lambda_t^\dagger\right)$.
4. $Y_{t+1}|\lambda_t^\dagger \sim \text{Binomial}\left(N, \lambda_t^\dagger\right)$.
5. $Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left(S_t\,(1 - e^{-\lambda_t^\dagger})\right)$.
6. $Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left(N\,(1 - e^{-\lambda_t^\dagger})\right)$.

7. $Y_{t+1} | \lambda_t^\dagger \sim \text{Poisson}\left(S_t\, \lambda_t^\dagger\right)$.

8. $Y_{t+1} | \lambda_t^\dagger \sim \text{Poisson}\left(N\, \lambda_t^\dagger\right)$.

For all eight models, the force of infection is modeled as $\lambda_t^\dagger = e^{\alpha_{\text{AR}}} Y_t / N + e^{\alpha_{\text{EN}}}$. In Figure 1, we plot the average parameter estimates from each of the eight models under the three values of R_0, along with the 2.5- and 97.5-percentiles of the estimates across simulations. In Figure 1(a), where $R_0 = 2.5$, the epidemic is limited by the number of susceptibles, and dies off when there are few remaining susceptible individuals in the population. In this setting, we see that those models that approximate the number of susceptibles with the initial number of susceptibles do not perform well. Although less dramatic, estimates from models that made the Taylor approximation of risk perform worse than those that do not make the approximation. However, with such explosive growth, there is limited variability in the simulated epidemics, and as a result the range of estimates of α_{AR} is so narrow that the intervals are undetectable in the upper panel of 1(a); further details of these results are included in the web material. In Figures 1(b) and 1(c), where $R_0 = 1$ and $R_0 = 0.85$, and the epidemic is not growing as dramatically, we see that the simplifying assumptions necessary for the ecological vaccine model are more appropriate. While there is some slight underestimation of the autoregressive term and overestimation of the endemic term due to the finite sample size, the estimated bias and MSE are similarly small for all eight models (see web material).

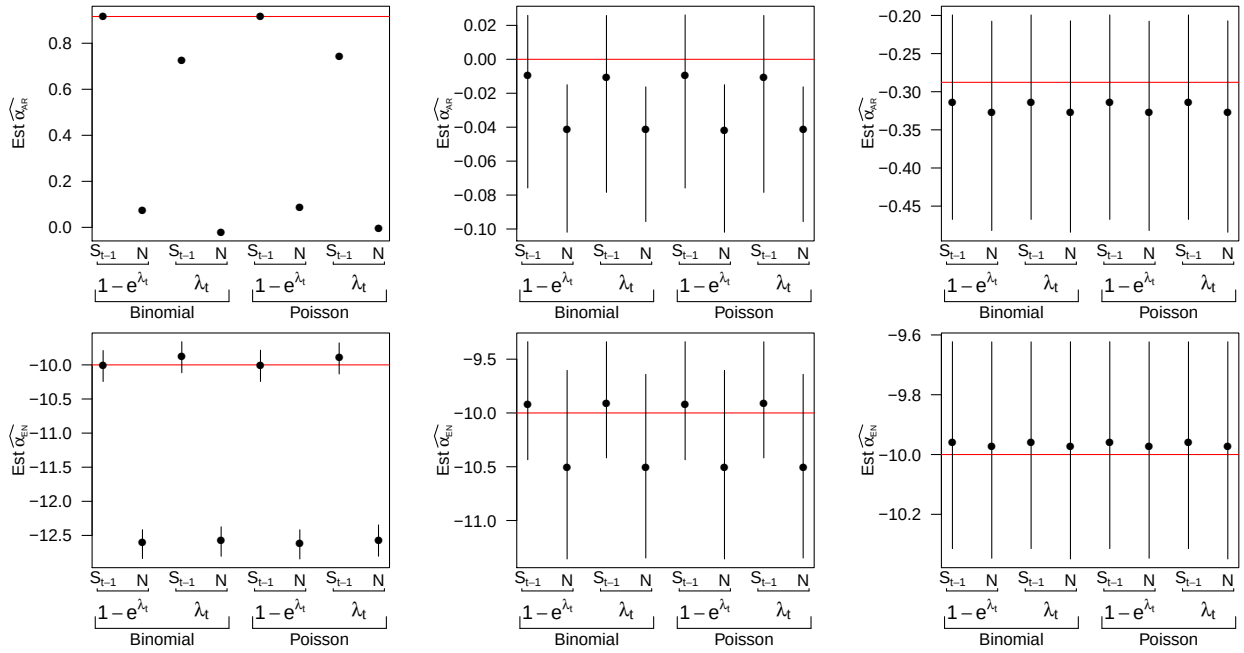
4.2 | Assessing the ecological model in a partially vaccinated population

We now consider the performance of the ecological model within a partially vaccinated population. For identifiablity, we consider $i = 5$ areas, each with $N_i = 100,000$ and that have varying levels of vaccine coverage. We focus on scenarios in which we expect the ecological vaccine model to perform well. The results from Section 4.1 showed that the ecological vaccine model performed well when $R_0 < 1$, which corresponds to $R < 1$ in a partially vaccinated population. Assuming $R_0 = 2.5$ and a vaccine effect of 0.8, we let vaccine coverages range from 65% to 85%. We simulate 250 epidemics assuming either an all-or-none vaccine or a leaky vaccine. Each simulated epidemic assumes a single infected individual who is unvaccinated to start, so that $Y_{ui0} = 1$ and $Y_{vi0} = 0$; and the initial number of susceptibles by vaccination status (S_{ui0} and S_{vi0}) is determined by the assumed vaccine mode of action (see Table 1). The number of cases by vaccination status are simulated as follows:

$$Y_{ui,t+1} \mid \lambda_{uit}^\dagger \sim \text{Binomial}\Big(S_{uit}, 1 - \exp(-\lambda_{uit}^\dagger)\Big), \tag{14}$$

$$Y_{vi,t+1} \mid \lambda_{vit}^\dagger \sim \text{Binomial}\Big(S_{vit}, 1 - \exp(-\lambda_{vit}^\dagger)\Big), \tag{15}$$

where the forms of $\lambda_{uit}^\dagger$ and $\lambda_{vit}^\dagger$ are determined by the assumed vaccine mode of action. The underlying force of infection is $\lambda_{it}^\dagger = \exp(\alpha_{\text{AR}})\,(Y_{uit} + Y_{vit})/N_i + \exp(\alpha_{\text{EN}})/N$, where $N = \sum_i N_i$. As in the previous simulations, we discard the first four time



(a) $R_0 = 2.5$ (b) $R_0 = 1$ (c) $R_0 = 0.85$

FIGURE 1 Summary of simulation results assessing simplifying assumptions. Average parameter estimates and intervals extending from the 2.5th and 97.5th percentile of estimates across simulations. Rows correspond to the parameter, columns to the true values of R_0. The first row shows estimates of α_{AR}; the second row depicts estimates of α_{EN}. True parameter values are denoted by red lines.

steps before simulating the equivalent of 3 years of weekly counts. We assume there are no infections from other areas, i.e. no neighborhood component. We compare MLE estimates obtained via numerical optimization from the following models:

1. Fully observed all-or-nothing model:

$$Y_{ui,t+1} \mid \lambda_{it}^{\dagger}, \phi \sim \text{Binomial}\left(S_{uit}(\phi), 1 - e^{-\lambda_{it}^{\dagger}}\right),$$
$$Y_{vi,t+1} \mid \lambda_{it}^{\dagger}, \phi \sim \text{Binomial}\left(S_{vit}(\phi), 1 - e^{-\lambda_{it}^{\dagger}}\right),$$
$$S_{uit}(\phi) = (1 - x_i)N_i - \sum_{k=1}^{t} Y_{uik},$$
$$S_{vit}(\phi) = (1 - \phi)x_i N_i - \sum_{k=1}^{t} Y_{vik}.$$

2. Fully observed leaky model:

$$Y_{ui,t+1} \mid \lambda_{it}^{\dagger}, \phi \sim \text{Binomial}\left(S_{uit}, 1 - e^{-\lambda_{it}^{\dagger}}\right),$$
$$Y_{vi,t+1} \mid \lambda_{it}^{\dagger}, \phi \sim \text{Binomial}\left(S_{vit}, 1 - e^{-(1-\phi)\lambda_{it}^{\dagger}}\right),$$
$$S_{uit} = (1 - x_i)N_i - \sum_{k=1}^{t} Y_{uik},$$
$$S_{vit} = x_i N_i - \sum_{k=1}^{t} Y_{vik}.$$

3. Ecological vaccine model:

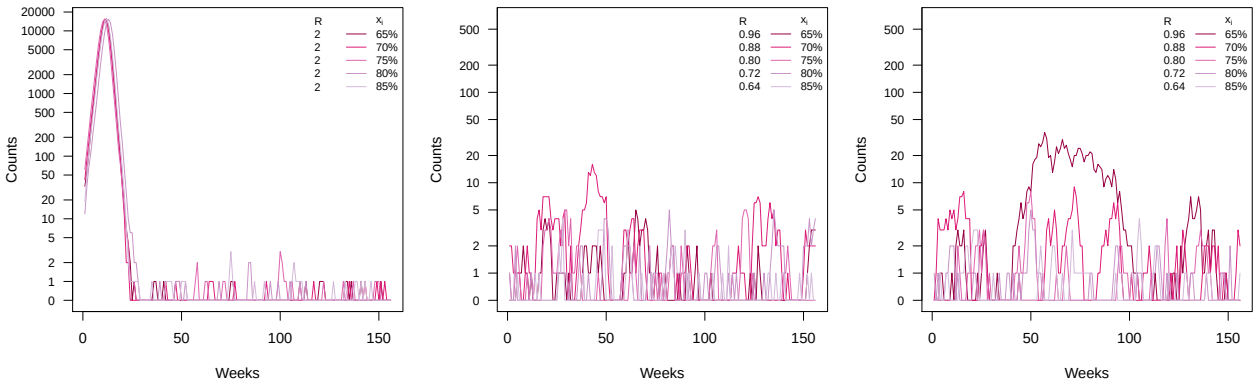
$$Y_{i,t+1} | \lambda_{it}^{\dagger} \sim \text{Poisson}\left(N_i(1 - \phi x_i)\lambda_{it}^{\dagger}\right).$$

4. Epidemic-endemic model:

$$Y_{i,t+1} | \mu_{it} \sim \text{Poisson}(\mu_{it}),$$
$$\mu_{it} = \exp(\alpha_0)\,(1 - x_i)^{\alpha_1}\, Y_{it} + \frac{N_i}{N} \exp(\beta_0),$$

We have parameterized $\lambda_{it}^{\dagger}$ in models 1–3 so that α_0 and β_0 in the epidemic-endemic model are comparable to α_{AR} and α_{EN}, respectively, in the other models. Note that the parameter associated with vaccine coverage in the epidemic-endemic model, α_1, is not directly comparable to the vaccine effect ϕ of the other models. Additionally, both the all-or-none (1) and the leaky (2) models assume that we have observed the number of cases by vaccination status, which is not necessary for the ecological (3) and epidemic-endemic (4) models.

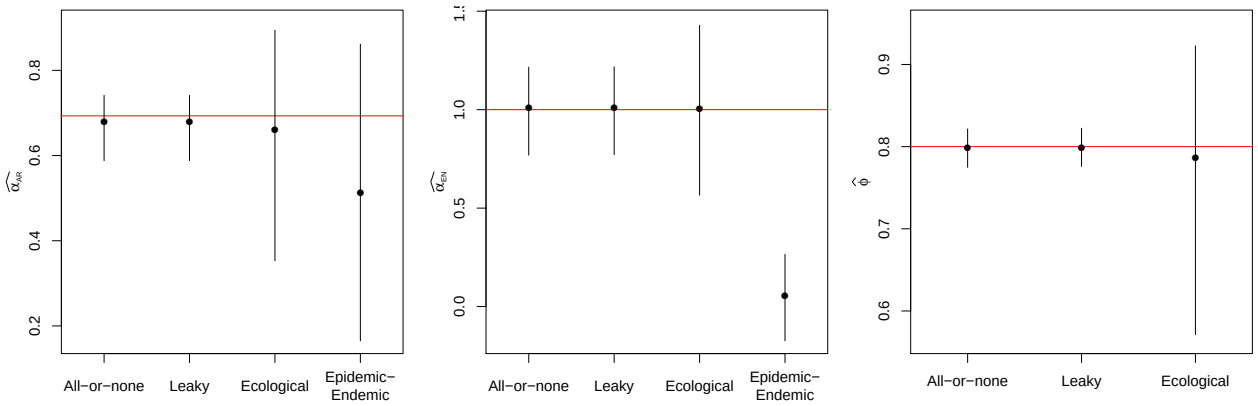
In Figure 2 we present an example of realizations for the five populations under the assumption of no vaccine effect, an all-or-none vaccine, and a leaky vaccine, with an assumed vaccine effect of $\phi = 0.8$.



(a) No vaccine effect ($\phi = 0$) (b) All-or-none vaccine ($\phi = 0.8$) (c) Leaky vaccine ($\phi = 0.8$)

FIGURE 2 Simulated epidemic curves for five populations, when there is (a) no vaccine effect, (b) an effective all-or-none vaccine, and (c) an effective leaky vaccine. Darker lines correspond to areas with lower vaccination coverage. Corresponding effective reproductive numbers (R) are included with vaccination coverage (x_i) in the legend.

In Figures 3 and 4 we present the average estimates, along with the 2.5- and 97.5-percentiles of estimates obtained under all four models, when the data were simulated assuming an all-or-none or leaky vaccine, respectively. Under all scenarios, the fully observed models yield estimates close to the true model parameters. Compared to the fully observed model estimates, the ecological vaccine model obtains similar estimates, but with wider intervals, appropriately reflecting the lost information as a result of the aggregation. In contrast, the epidemic-endemic models yield estimates that are very different from the true autoregressive and endemic parameter values. We do not include the epidemic-endemic estimates in the pictures for the estimates of the vaccine effect, ϕ, since the epidemic-endemic parameter is not comparable to the parameters in the other models.



(a) $\widehat{\alpha_{\text{AR}}}$ (b) $\widehat{\alpha_{\text{EN}}}$ (c) $\hat{\phi}$

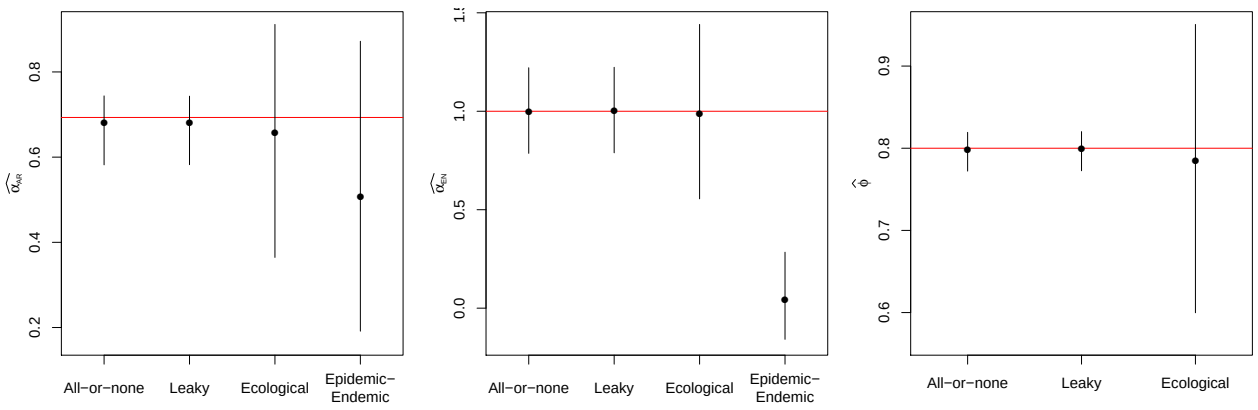
FIGURE 3 Summary of simulation results of partially vaccinated populations, assuming an *all-or-none* vaccine. Average estimates and intervals extending from the 2.5th and 97.5th percentile of estimates across simulations of (a) α_{AR}, (b) α_{EN}, and (c) ϕ for the fully observed all-or-none and leaky models, the ecological vaccine model, and the epidemic-endemic model. Red horizontal lines denote the true parameter values.

These simulations also provide a clear example of the risk for ecological bias when using the epidemic-endemic model. Interpreting the results from the epidemic-endemic model as individual-level parameter estimates would result in erroneous conclusions, especially regarding the endemic risk.

We also consider the results from 20 years worth of data in Appendix C and see that asymptotically, the ecological vaccine model yields unbiased estimates for all model parameters, consistent with the fully observed models.

5 | APPLICATION TO MEASLES DATA

We now apply the ecological vaccine model to data collected on measles outbreaks in Germany from 2005 through 2007. Measles is a highly contagious viral infection that can result in death for young or malnourished children. The average number of secondary infections that arise from a single measles infection in a completely susceptible population is estimated to be



(a) $\widehat{\alpha_{\text{AR}}}$ (b) $\widehat{\alpha_{\text{EN}}}$ (c) $\hat{\phi}$

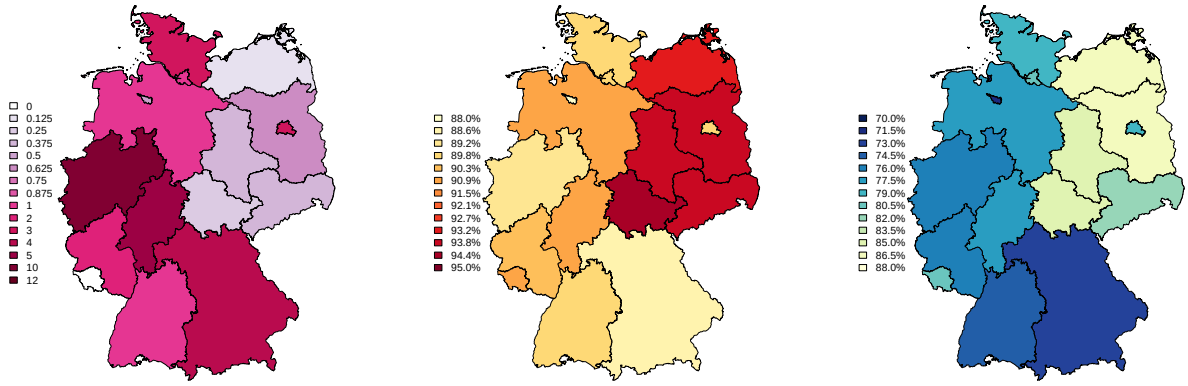
FIGURE 4 Summary of simulation results of partially vaccinated populations, assuming an *leaky* vaccine. Average estimates and intervals extending from the 2.5th and 97.5th percentile of estimates across simulations of (a) α_{AR}, (b) α_{EN}, and (c) ϕ for the fully observed all-or-none and leaky models, the ecological vaccine model, and the epidemic-endemic model. Red horizontal lines denote the true parameter values.

between 15 and 20.[21,22] Fortunately, the measles, mumps, and rubella (MMR) vaccine is very effective. Between 85% and 95% of children will develop immunity after a single dose of the MMR vaccine and a second dose provides nearly 99% vaccine efficacy.[22] Even with an effective vaccine the highly infectious nature of measles means that more than 93% of the population needs to be immune in order to prevent epidemics.[23,24] Hence, even in countries with well establish vaccination programs, such as Germany, small outbreaks persist.

We use data from Germany's national disease surveillance system which has been previous used to examine the relationship between vaccination coverage and the size of measles outbreaks and is included in the `surveillance` package for `R`.[9] Further details about this data and previous analysis can be found elsewhere.[12] For our analyses, we assume a two week time step, based on the approximate generation time for measles.[12,25] Between 2005 and 2007, over 3,500 cases of measles were reported throughout Germany, with as many as 344 cases observed in a single bi-week. Over the three years, no cases were observed in Saarland, and approximately 2,000 of those cases were observed in the state of North Rhine-Westphalia, see Figure 5(a).

Estimated MMR vaccination coverage is based on the number of students presenting vaccination cards at the required medical exam for school entry.[12] Between 87% and 95% of students brought vaccination cards to the entry exam preceding the start of the 2006–2007 school year. Following the previous analysis, we estimate the coverage for at least one MMR vaccine by assuming that the coverage in the population that did not bring the vaccination cards is half that of those who did have vaccination cards.[12] In Figures 5(b) and 5(c), we map the estimated vaccine coverage for one or more MMR vaccines (left) and at least two vaccines (right). Although the available vaccination data is for children starting primary school, typically between 4 and 7 years of age, we assume that the MMR vaccination coverage for the whole population is the same as the estimated vaccination coverage for this analysis. We note that the estimated coverage is likely to be an overestimate, as those who show up for the annual medical

exam and bring vaccination cards are more likely to have more complete medical records.[12] We summarize the number of cases



(a) Cases per 100,000. (b) At least 1 MMR vaccination. (c) At least 2 MMR vaccinations.

FIGURE 5 Total number of measles cases per 100,000 observed between 2005 and 2007 (a). Estimated vaccine coverage for at least 1 MMR vaccination (b) and at least 2 MMR vaccinations (c) in 2006, based on data from examination of vaccination cards in school aged children.

and estimated coverage in Table D1.

In this analysis, we are primarily interested in estimating the effects of vaccination on the observed cases of measles. We expand the ecological vaccine model developed in previous sections to incorporate spatial and temporal dependencies. In addition, we adopt a Bayesian paradigm, in order to incorporate our previous knowledge about the MMR vaccine effectiveness. We fit the following ecological model to the measles data:

$$
\begin{aligned}
Y_{i,t+1} \mid \mu_{it}, \phi &\sim \text{Poisson}\left(N_i(1-\phi x_i)\left(\lambda_i \frac{y_{it}}{N_i} + \nu_{it}\right)\right), \\
\log \lambda_i &= \alpha_{\text{AR}} + a_i, \\
\log \nu_{it} &= \alpha_{\text{EN}} + b_i + \gamma \sin(\omega_t) + \delta \cos(\omega_t) - \log(N), \\
a_i &\sim \text{N}(0, \sigma^2_{\text{AR}}), \\
b_i &\sim \text{N}(0, \sigma^2_{\text{EN}}), \\
\phi &\sim \text{Beta}(10, 2.5),
\end{aligned}
\tag{16}
$$

where x_i is the estimated vaccine coverage in area i; component-specific random effects a_i and b_i are assumed independent; $\omega_t = 2\pi t/26$; and the beta prior on ϕ places 90% of the mass is between 0.6 and 0.99. We assume lognormal priors with large variances on α_{AR} and α_{EN}. In the formulation of λ_i, we have assumed transmission to be frequency dependent based on previous studies of measles in England and Wales.[25] Hamiltonian Monte Carlo (HMC) sampling via `Stan` was used to fit this more complex ecological model.[26] Corresponding code can be found in the web material.

In Table 2, we summarize the posterior estimates of the fixed effects from the ecological vaccine model. We estimate the vaccine effect to be 0.92, with a 95% posterior credible interval from 0.66 to 0.99, which is commensurate with the known vaccine efficacy for the MMR vaccine. However, this estimate is also similar to the strong prior placed on ϕ (prior 95% interval is from 0.55 to 0.96). Vaccine coverage ranges from 88% to 95% across the 16 German states, and this results suggests that there is little information about the vaccine effect in these data. As a sensitivity analysis, we fit the same hierarchical model with a non-informative prior for ϕ. The results are not presented here, but can be found in the web material. The non-informative prior on ϕ results in slightly higher estimates for both α_{AR} and ϕ, but each have substantially wider credible intervals. The prior choice for ϕ had little effect on the posterior estimates of the parameters in the endemic component of the model.

	Median	2.5%	97.5%
α_{AR}	0.91	-0.26	1.66
ϕ	0.92	0.66	0.99
α_{EN}	3.53	2.54	4.16
γ	0.71	0.55	0.86
δ	-0.20	-0.36	-0.04
σ_{AR}	0.66	0.28	1.61
σ_{EN}	0.52	0.28	0.96
R_0	2.49	0.77	5.24

TABLE 2 Posterior medians and 95% credible intervals for the parameters of the ecological model for the measles data.

We plot the posterior median and 95% credible intervals for state-specific autoregressive parameters from the ecological vaccine model, computed as $(1-\hat{\phi}x_i)\ \exp(\widehat{\alpha_{\text{AR}}}+\hat{a}_i)$ in Figure 6. Notice that for the ecological vaccine model, the autoregressive parameter has an intuitive interpretation as the effective reproductive number, where $R = (1 - x\phi)R_0$.[19] As expected, all estimates were below 1, but the area-specific estimates have credible intervals with varying widths. The widest interval was observed for Saarland, and the smallest for North Rhine-Westphalia, the two states with the fewest (0) and most (2,036) observed cases over the three year study.

In Figure 7, we plot the total number of observed measles cases and prevalence per 100,000 people, by state and biweek for the sixteen states in Germany. The left axis indicates the total number of cases; the right axis indicates the prevalence per 100,000 people. The estimated vaccine coverage and effective reproductive number are included the upper left and right corners of each frame. Fitted values are included in the red and computed following (16) as

$$\hat{Y}_{it} = (1-\hat{\phi}x_i)\left[\exp\left(\widehat{\alpha_{\text{AR}}}+\hat{a}_i\right)\ Y_{i,t-1} + (N_i/N)\ \exp\left(\widehat{\alpha_{\text{EN}}}+\hat{b}_i+\hat{\gamma}\sin(\omega_t)+\hat{\delta}\cos(\omega_t)\right)\right], \tag{17}$$

where $Y_{i,t-1}$ is the observed number of counts for area i and week $t-1$, and $\omega_t = 2\pi t/26$. In general, the ecological vaccine model provides good estimates for the number of cases.

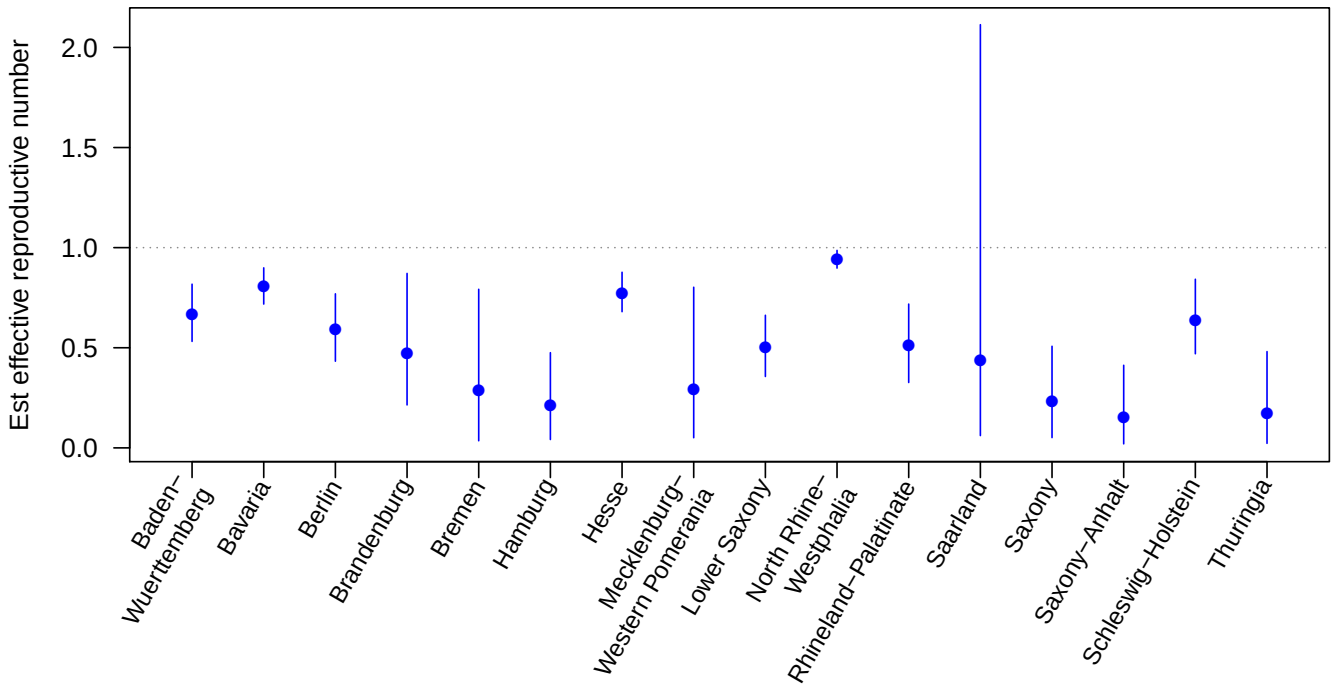


FIGURE 6 Estimated state-level autoregressive components. For the ecological vaccine model, posterior median and 95% credible intervals are presented from the Stan model fit.

In Figure D3 we plot the area-specific random effects for the autoregressive and endemic components. The states with the highest prevalence have higher autoregressive random effects. The endemic random effects do not appear to have a similar spatial structure as the autoregressive random effects. Moreover, when the autoregressive random effects are plotted against the endemic random effects, as in Figure D4, there is no evidence of a strong correlation between the two components. This supports our decision to model the component-specific random effects as independent. However, in other settings, we may want to consider more complex forms of random effects. For example, if there were strong correlations between the component-specific random effects, it may be more appropriate to assume bivariate normal distribution for the random effects.

In this analysis, the posterior estimate of R_0 is 2.49 (95% CI: 0.77 – 5.24), which is much smaller than the typical R_0 between 15 and 20 for measles.[21,22] There are many possible sources of this underestimation. Our analyses (and the available data) are in discrete time (bi-weeks), but in reality, new infections occur in continuous time and space. The discretization of time is known to result in a biased estimate of R_0.[27] It is likely that large outbreaks, like that in North Rhine-Westphalia in 2006 prompted additional vaccination campaigns. However, we have only a single estimate of vaccination coverage, from children entering school. The estimation of vaccine coverage is likely to not capture the true levels of protection within the population, or the heterogeneity of protection across various age groups.[28] Lastly, with any disease surveillance system there is likely to be under-reporting of cases. One study of a single German state found that under-reporting varied dramatically over the course of the outbreak.[29]

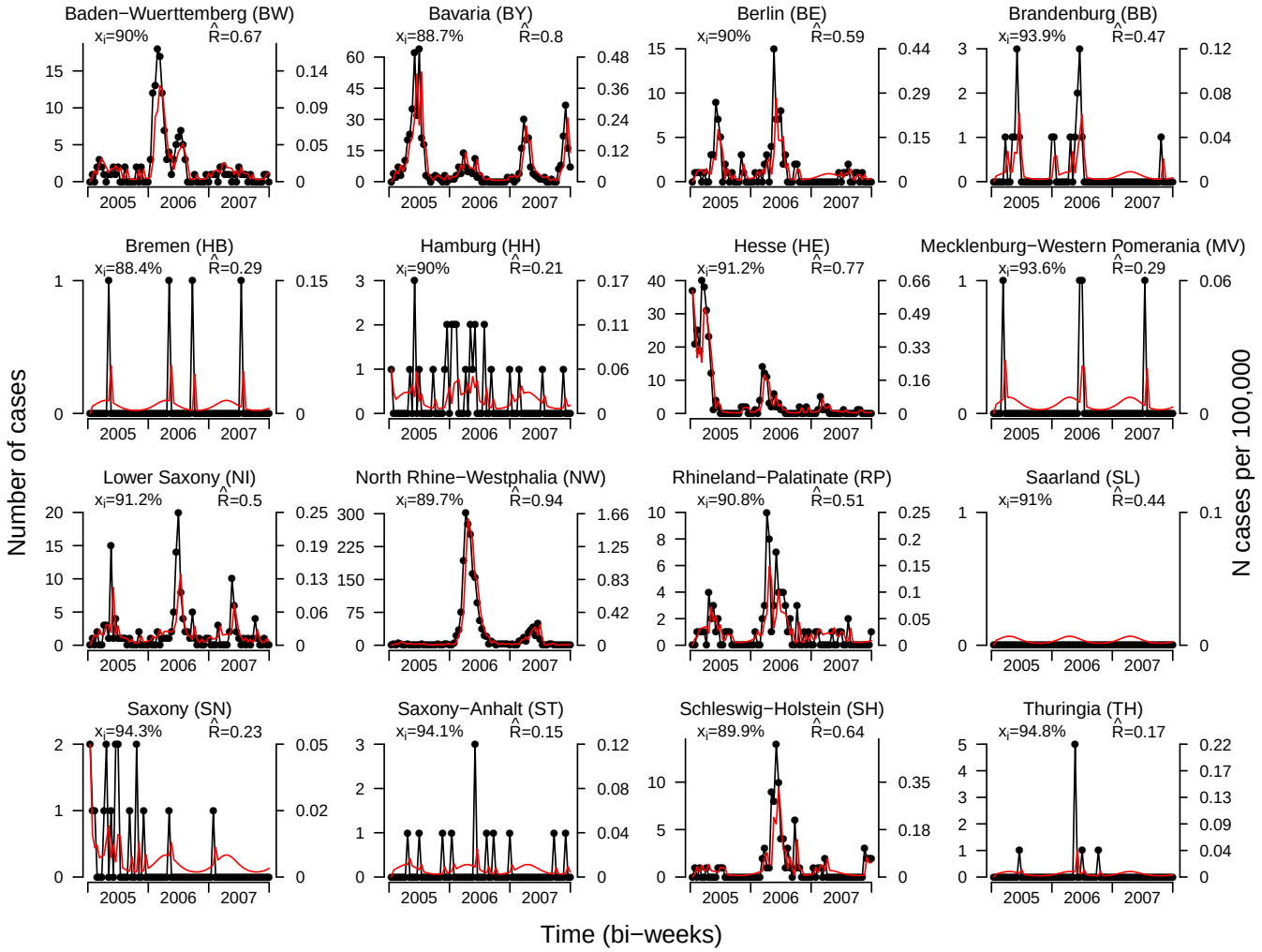


FIGURE 7 Number of measles cases and prevalence by state and biweek from 2005 through 2007. The left axis indicates the total number of cases; the right axis indicates the prevalence per 100,000 people. Estimated MMR vaccine coverage is included in the upper left corner of each plot. Fitted values from the ecological vaccine model is included in red. Estimated effective reproductive numbers ($\hat{R}$) is included in the upper right corner.

6 | DISCUSSION

Infectious disease surveillance data is the primary source of information about disease spread in large populations over time. Current approaches to analyzing these sorts of data tend to focus on prediction, but when used to study covariate effects, the parameter interpretation is cumbersome, especially when the interest is in understanding how vaccination coverage associates with disease. With inference in mind, we started with an individual-level model that included how vaccination affect risk of infection and derived an ecologically consistent model for infectious disease data that accounts for vaccination coverage. A key benefit to our approach is that we obtain estimates of familiar epidemiological parameters, which are easy to interpret (though caveats are in order due to other issues, see the discussion at the end of Section 5). And we saw that under common simplifying assumptions, the resulting ecological vaccine model is the same regardless of the assumed mode of vaccine action. Simulations

showed that the ecological vaccine model performs reasonably well in many practical scenarios and illuminated situations when the ecological vaccine model may be inappropriate.

There are limitations to the current model, and important extensions to make the approach more broadly applicable. For example, it would be beneficial to extend the ecological vaccine model to account for a non-constant and perhaps longer infectiousness period. It may be interesting to consider bivariate random effects, or spatially structured random effects in the autoregressive and/or endemic components. Future work will be focused on extending the method to account for stratified population structures and including neighborhood effects in the ecological vaccine model.

`Stan` and `R` code to fit the models of this paper can be found in the supporting information for this article at https://github.com/lhfisher/Ecological_Inference.

ACKNOWLEDGMENTS

The authors would like to thank Elizabeth Halloran and Jonathan Sugimoto for helpful suggestions on an earlier draft of this manuscript. We would also like to thank the reviewers for their helpful feedback.

Author contributions

LHF developed the model, ran simulations, and drafted the manuscript; JW developed the model and supervised the work.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:

Code for simulations. The code to set up, run, and summarize the various simulations presented in this manuscript is available at https://github.com/lhfisher/Ecological_Inference.

Code for the data analyses. The data and code for the measles data analysis presented in this manuscript is available at https://github.com/lhfisher/Ecological_Inference.

References

1. Selvin HC. Durkheim's 'suicide' and problems of empirical research. *American Journal of Sociology* 1958; 63: 607–619.

2. Robinson WS. Ecological correlations and the behavior of individuals. *American Sociological Review* 1950; 15: 351–357.

3. Greenland S. Divergent biases in ecologic and individual level studies.. *Statistics in Medicine* 1992; 11: 1209–1223.

4. Greenland S, Robins J. Ecological studies: biases, misconceptions and counterexamples. *American Journal of Epidemiology* 1994; 139: 747–760.

5. Richardson S, Monfort C. Ecological Correlation Studies. In: Elliott P, Wakefield JC, Best NG, Briggs DJ., eds. *Spatial Epidemiology: Methods and Application*Oxford University Press, Oxford. 2000.

6. Wakefield J. Ecologic studies revisited. *Annual Review of Public Health* 2008; 29: 75–90.

7. Wakefield J, Lyons H. Spatial aggregation and the ecological fallacy. In: Gelfand A, Diggle P, Guttorp P, Fuentes M., eds. *Handbook of Spatial Statistics*CRC Press. 2010.

8. Finkenstädt BF, Grenfell BT. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2000; 49(2): 187–205.

9. Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 2005; 5: 187–199.

10. Paul M, Held L, Toschke AM. Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* 2008; 27: 6250–6267.

11. Paul M, Held L. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine* 2011; 30: 1118–1136.

12. Herzog SA, Paul M, Held L. Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. *Epidemiology and Infection* 2011; 139: 505–515.

13. Meyer S, Held L. Power-law models for infectious disease spread. *The Annals of Applied Statistics* 2014; 8: 1612–1639. doi: 10.1214/14-AOAS743

14. Meyer S, Held L, Höhle M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. *Journal of Statistical Software* 2017; 77(11). doi: 10.18637/jss.v077.i11

15. Bauer C, Wakefield J. Stratified space-time infectious disease modeling: with an application to hand, foot and mouth disease in China. *Journal of the Royal Statistical Society, Series A* 2018; 67: 1379–1398.

16. Xia Y, Bjørnstad ON, Grenfell BT. Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist* 2004; 164: 267–281.

17. Wakefield J, Dong T, Minin V. Spatio-Temporal Analysis of Surveillance Data. In: Gelfand A, Diggle P, Guttorp P, Fuentes M., eds. *Handbook of Spatial Statistics*CRC Press. 2019.

18. Koopman JS, Longini IM. The ecological effects of individual exposures and nonlinear disease dynamics in populations.. *American Journal of Public Health* 1994; 84: 836–842. doi: 10.2105/AJPH.84.5.836

19. Halloran ME, I.M. Longini J, Struchiner CJ. *Design and Analysis of Vaccine Studies*. New York: Springer . 2010.

20. Van Boeckel TP, Takahashi S, Liao Q, et al. Hand, Foot, and Mouth Disease in China: Critical Community Size and Spatial Vaccination Strategies. *Scientific Reports* 2016; 6(1): 25248. doi: 10.1038/srep25248

21. Keeling MJ, Rohani P. *Modeling Infectious Diseases in Humans and Animals*. Princeton and Oxford: Princeton University Press . 2008.

22. Sudfeld CR, Navar AM, Halsey NA. Effectiveness of measles vaccination and vitamin A treatment. *International Journal of Epidemiology* 2010; 39: i48–i55. doi: 10.1093/ije/dyq021

23. Centers for Disease Control and Prevention . Measles, Mumps, and Rubella – Vaccine Use and Strategies for Elimination of Measles, Rubella, and Congenital Rubella Syndrome and Control of Mumps: Recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR* 1998; 47(No. RR-8): 1–58.

24. World Health Organization . Measles vaccines: WHO position paper. *Weekly Epidemiological Record* 2009; 84: 349–360.

25. Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* 2002; 72: 169–184.

26. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 2017; 76(1). doi: 10.18637/jss.v076.i01

27. Ferrari MJ, Bjørnstad ON, Dobson AP. Estimation and inference of R0 of an infectious pathogen by a removal method. *Mathematical Biosciences* 2005; 198(1): 14–26. doi: 10.1016/j.mbs.2005.08.002

28. Poethko-Müller C, Mankertz A. Seroprevalence of measles-, mumps- and rubella-specific IgG antibodies in German children and adolescents and predictors for seronegativity. *PLoS ONE* 2012; 7(8): 1–13. doi: 10.1371/journal.pone.0042867

29. Mette A, Reuss AM, Feig M, et al. Under-reporting of measles–an evaluation based on data from North Rhine-Westphalia. *Deutsches Arzteblatt International* 2011; 108(12): 191–6. doi: 10.3238/arztebl.2011.0191

30. Wakefield J, Haneuse S, Dobra A, Teeple E. Bayes computation for ecological inference. *Statistics in medicine* 2011; 30: 1381–1396.

31. Richardson S, Stucker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology* 1987; 16: 111–120.

32. Plummer M, Clayton D. Estimation of population exposure. *Journal of the Royal Statistical Society, Series B* 1996; 58: 113–126.

How to cite this article: Ecological inference for infectious disease data, with application to vaccination strategies, *Statistics in Medicine*, *2018;00:1–6*.

APPENDIX

A ECOLOGICAL BIAS FOR INFECTIOUS DISEASE MODELS

To better understand ecological bias in the infectious disease setting, we start with a simple individual-level model. Recall, ecological bias occurs when a naïve ecological model is used to make conclusions on individual-level parameters but the implied aggregate risk differs from that of the individual. Let Y_{itj} be the disease indicator for susceptible individual j in week t and area i, where $j = 1, \dots, n_i$. Assuming a rare disease so that $1 - exp(\lambda^{\dagger}_{itj}) \approx \lambda^{\dagger}_{itj}$, we start with the individual-level model

$$Y_{itj}|y_{i,t-1} \sim \text{Bernoulli}\left(\lambda^{\dagger}_{itj}\right), \tag{A1}$$

where $\lambda^{\dagger}_{itj} = \lambda^{\text{AR}}_{itj}\, y_{i,t-1}/n_i + \lambda^{\text{EN}}_{itj}$ and $\lambda^{\text{AR}}_{itj}$ and $\lambda^{\text{EN}}_{itj}$ are individual-level risks. We additionally assume that the individual risk of infection is a function of some individual-level covariates z_{itj} such that

$$\begin{aligned} \lambda^{\text{AR}}_{itj} &= e^{\alpha_0} f(\alpha_1, z_{itj}), \\ \lambda^{\text{EN}}_{itj} &= \lambda^{\text{EN}}_{it}, \end{aligned} \tag{A2}$$

where $f(\alpha_1, z)$ describes the relationships between the covariate and component-specific risk. In the rare disease setting, the aggregate model for the total number of cases in area i and time t implied by the individual-level model in (A1) and (A2) is

$$Y_{it}|y_{i,t-1} \sim \text{Poisson}\left(\overline{\lambda^{\text{AR}}_{it}}\, y_{i,t-1}/n_i + \overline{\lambda^{\text{EN}}_{it}}\right), \tag{A3}$$

where $\overline{\lambda_{it}^{\mathrm{AR}}}$ and $\overline{\lambda_{it}^{\mathrm{EN}}}$ are the aggregate autoregressive and endemic risks. We have assumed a constant endemic risk and therefore, $\overline{\lambda_{it}^{\mathrm{EN}}} = n_i \lambda_{it}^{\mathrm{EN}}$. The form of the aggregate autoregressive risk, $\overline{\lambda_{it}^{\mathrm{AR}}}$, will depend on the form of the covariate. For a continuous covariate, the autoregressive aggregate risk is

$$\overline{\lambda_{it}^{\mathrm{AR}}} = e^{\alpha_0} \int_{A_i} f(\alpha_1, z)\, g_{it}(z|\boldsymbol{\omega}_{it}) dz, \tag{A4}$$

where z is assumed to be distributed $g_{it}(z|\boldsymbol{\omega}_{it})$, with area- and week-level parameters for that distribution $\boldsymbol{\omega}_{it}$; and where A_i represents area i. For a discrete individual-level covariate, z_k with K levels, the aggregate risk implied by the individual-level model is

$$\overline{\lambda_{it}^{\mathrm{AR}}} = e^{\alpha_0} \sum_{k=1}^{K} f(\alpha_1, z_k)\, g_{it}(z_k|\boldsymbol{\omega}_{it}). \tag{A5}$$

In other words, the consistent aggregated risk is found by averaging the individual-level risk over the distribution of the covariate within area i and week t.

However, when only the aggregated data is available, analyses are limited to modeling total number of cases $Y_{it} = \sum_{j=1}^{n_i} Y_{itj}$, and the area- and week-specific average exposures, $\bar{z}_{it}$.It is tempting to fit the naïve ecological regression model

$$\mathrm{E}\left[Y_{it}|y_{i,t-1}\right] = \exp(\beta_0 + \beta_1 \bar{z}_{it})\, y_{i,t-1} + \exp(\beta_2), \tag{A6}$$

where $\exp(\beta_1)$ is the relative risk of within-area infection associated with a one unit increase in the average exposure, $\bar{z}_{it}$. Therefore, the naïve ecological model assumes the aggregate risk is consistent with the individual-level risk, $\overline{\lambda_{it}^{\mathrm{AR}}} = \exp(\beta_0 + \beta_1 \bar{z}_{it})$.

Typically, the parameter estimates from (A6) will not be equal to those from implied aggregate model of (A3). The specific form of the implied aggregate risk will, therefore, depend on the within-area distribution of that specific covariate. For example, if $f(\alpha_1, z) = \exp(\alpha_1 z)$ and we assume the within-area exposures are distributed normally, i.e. $z \mid \bar{z}_{it},\ \sigma_{it}^2 \sim \mathrm{Normal}\left(\bar{z}_{it}, \sigma_{it}^2\right)$, the aggregate risk is

$$\overline{\lambda_{it}^{\mathrm{AR}}} = \exp\left(\alpha_0 + \alpha_1 \bar{z}_{it} + \alpha_1^2 \sigma_{it}^2/2\right). \tag{A7}$$

Thus the consistent aggregate risk is a function of both the average exposure and the variability of that exposure within a given area. Notice that when either the mean and variance are independent or when there is no within-area variability of exposures, $\sigma_{it}^2 = 0$ for all areas i and weeks t, the naïve model (A6) is identical to the consistent aggregate model (A7). For further details in a non-infectious disease setting see.[31,32] When the exposure is binary, implied aggregate risk is

$$\overline{\lambda_{it}^{\mathrm{AR}}} = e^{\alpha_0} \left[\left(1 - \bar{z}_{it}\right) + \bar{z}_{it}\, e^{\alpha_1}\right], \tag{A8}$$

where $\bar{z}_{it}$ is the proportion of exposed individuals in area i and week t.

In the non-infectious disease setting, it is well understood that when data are aggregated to the group level, individual-level associations can become distorted, leading to ecological bias. In some ways, it is misleading to refer to this difference as bias. Both the implied aggregate and naïve model will produce unbiased estimates *of different parameters*. The naïve model estimates the risk associated with the average exposure, while the implied aggregate model estimates the average of individual risks.[30] The 'bias' comes from trying to estimate individual-level associations from a model that estimates average parameters.

B ECOLOGICAL VACCINE MODEL IDENTIFIABILITY

We derive the ecological vaccine model when the vaccine's mode of action is a combination of both leaky and all-or-none. Following the development in Section 3, we assume that a vaccine fails with probability θ and when it takes, reduces risk of infection by ϕ. Individuals will fit into one of three groups: unvaccinated, failed vaccinated, and vaccinated subscripted by u, f, and v respectively. Let x be the proportion of vaccinated individuals in a fully susceptible population of size N. Hence the initial susceptible population will be

$$S_{u0} = (1-x)N, \qquad S_{f0} = (1-\theta)xN, \qquad S_{v0} = \theta xN.$$

The force of infection for each group is defined as

$$\lambda^{\dagger}_{ut} = \lambda^{\dagger}_{t}, \qquad \lambda^{\dagger}_{ft} = \lambda^{\dagger}_{t}, \qquad \lambda^{\dagger}_{vt} = (1-\phi)\lambda^{\dagger}_{t}.$$

Together, these define disease progression

$$\begin{aligned}
Y_{u,t+1}|\lambda^{\dagger}_{ut} &\sim \text{Binomial}\left(S_{ut}, 1-e^{-\lambda^{\dagger}_{t}}\right),\\
Y_{f,t+1}|\lambda^{\dagger}_{ut} &\sim \text{Binomial}\left(S_{ft}, 1-e^{-\lambda^{\dagger}_{t}}\right),\\
Y_{v,t+1}|\lambda^{\dagger}_{ut} &\sim \text{Binomial}\left(S_{vt}, 1-e^{-(1-\phi)\lambda^{\dagger}_{t}}\right),
\end{aligned}$$

where $S_{gt} = S_{g0} - \sum_{s=1}^{t} Y_{gs}$, for $g = \{u, f, v\}$. And the aggregated model is a convolution of the binomials (and unvaccinated and failed vaccinated groups can be combined into a single group). When the binomial distributions can be approximated by Poisson distributions, this implies

$$Y_{t+1}|\lambda^{\dagger}_{t} \sim \text{Poisson}\left((S_{ut}+S_{ft})(1-e^{-\lambda^{\dagger}_{t}}) + S_{vt}1-e^{-(1-\phi)\lambda^{\dagger}_{t}}\right).$$

The Taylor approximation simplifies the above

$$Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left((S_{ut} + S_{ft} + (1-\phi)S_{vt})\lambda_t^\dagger\right).$$

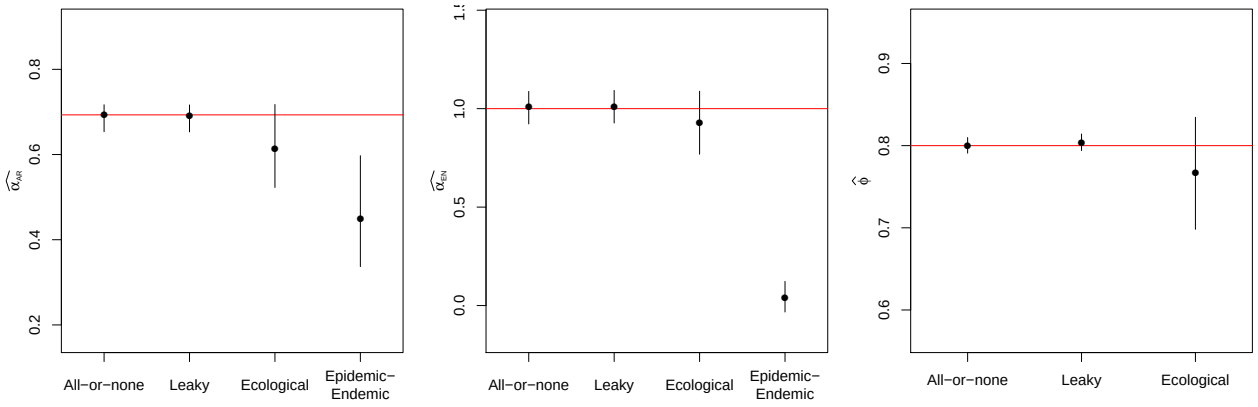
And when the number of infections is negligible, so that the number of susceptibles is approximately the initially susceptible population, we arrive at the ecological vaccine model

$$Y_{t+1}|\lambda_t^\dagger \sim \text{Poisson}\left((1-\phi\theta x)N\lambda_t^\dagger\right). \tag{B9}$$

Notice that in the above, the specific modes of vaccine action (all-or-none or leaky) cannot be identified with aggregate data.

C ASYMPTOTIC BEHAVIOR OF THE ECOLOGICAL VACCINE MODEL

Under the same conditions as the simulations in Section 4.2, we considered the results for ten years worth of data. In Figure C1 we present estimates from the fully observed all-or-none and leaky models, along with estimates from the ecological vaccine model and the epidemic-endemic model. We see that the estimates for the fully observed models, as well as the ecological vaccine models are much closer to the true parameter values compared to the previous simulations, which used only 3 years of weekly data. With long time series, the ecological vaccine model provides unbiased estimates for all model parameters.



(a) $\widehat{\alpha_{\text{AR}}}$ (b) $\widehat{\alpha_{\text{EN}}}$ (c) $\hat{\phi}$

FIGURE C1 Estimates and 95% confidence intervals for the fully observed all-or-none and leaky models, the ecological vaccine model, and the epidemic-endemic model for 10 years worth of weekly data simulated assuming an *all-or-none* vaccine. Red horizontal lines denote the true parameter values.

D ADDITIONAL RESULTS FROM MEASLES ANALYSIS

Table D1 presents the total number of measles cases and the estimates of vaccination coverage for the 16 states of Germany. In Figure D2 we plot a histogram of posterior samples of $\hat{\phi}$ along with the prior Beta(10, 2.5) curve. The posterior is similar

			Est. Coverage (%)	
State (Abbreviation)	Population	Total Cases	1st dose	2nd dose
Baden-Wuerttemberg (BW)	10,738,753	162	90.0%	75.6%
Bavaria (BY)	12,492,658	606	88.7%	73.2%
Berlin (BE)	3,404,037	104	90.0%	80.2%
Brandenburg (BB)	2,547,772	18	93.9%	86.9%
Bremen (HB)	663,979	4	88.4%	71.9%
Hamburg (HH)	1,754,182	29	90.0%	80.5%
Hesse (HE)	6,075,359	336	91.2%	78.1%
Mecklenburg-Western Pomerania (MV)	1,693,754	4	93.6%	88.0%
Lower Saxony (NI)	7,982,685	144	91.2%	78.0%
North Rhine-Westphalia (NW)	18,028,745	2,036	89.7%	76.9%
Rhineland-Palatinate (RP)	4,052,860	85	90.8%	77.3%
Saarland (SL)	1,043,167	0	91.0%	81.8%
Saxony (SN)	4,249,774	18	94.3%	82.4%
Saxony-Anhalt (ST)	2,441,787	12	94.1%	86.5%
Schleswig-Holstein (SH)	2,834,254	89	89.9%	79.3%
Thuringia (TH)	2,311,140	8	94.8%	85.9%

TABLE D1 Number of measles cases and estimated vaccination coverage for the 16 German states from 2005-2007. Estimated vaccination coverage for at least 1 or 2 MMR vaccinations and comes from the school entry examinations. Note this partially reproduces Table 1 from previous analyses [12].

to the prior, suggesting that there is little information about the vaccine effect in this data. As a sensitivity analysis, we fit the same hierarchical model with a non-informative prior for ϕ. The results for this analysis are presented in Appendix D. The non-informative prior on ϕ results in slightly higher estimates for both α_{AR} and ϕ, but each have substantially wider credible intervals. The prior choice for ϕ had little effect on the posterior estimates of the parameters in the endemic component of the model.

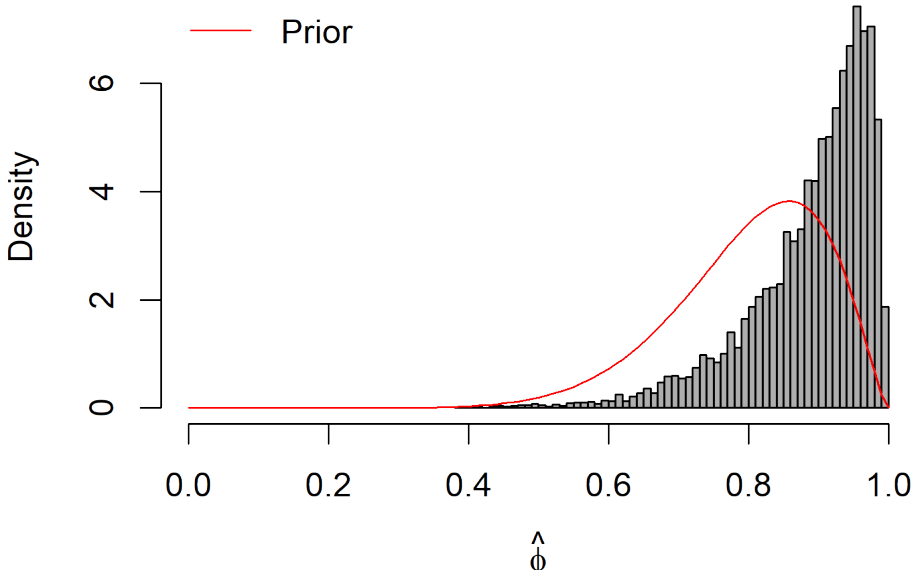
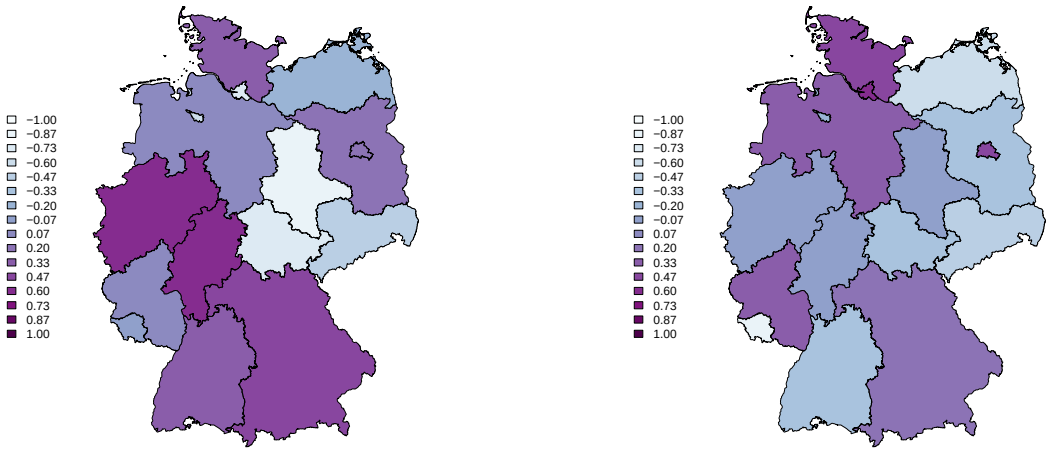


FIGURE D2 Histogram of posterior samples of $\hat{\phi}$. The red curve is the prior distribution, Beta(10, 2.5).



(a) AR random effects.

(b) EN random effects.

FIGURE D3 Maps of the random effect estimates for the autoregressive and endemic components in the ecological vaccine model.

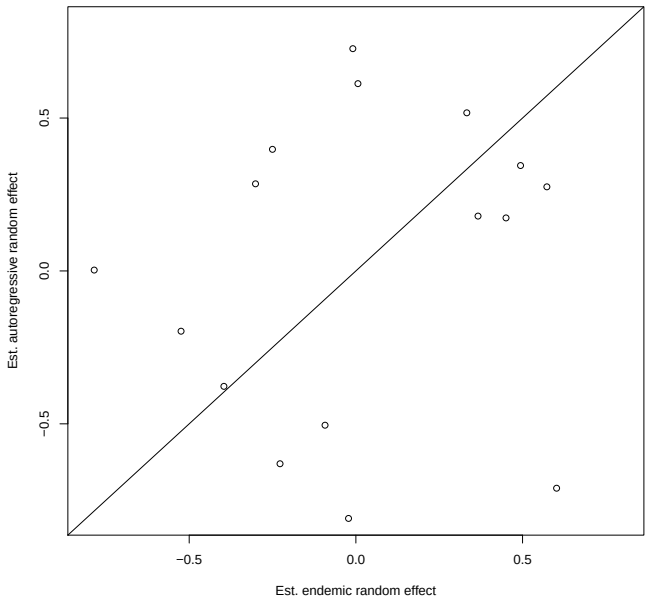


FIGURE D4 Comparison of autoregressive and endemic random effect estimates.