Inverse Problem Methods as a Public Health Tool in Pneumococcal Vaccination

Karyn L. Sutton[1,2] *, H. T. Banks[1], Carlos Castillo-Chavez[2,3,4]

[1] Center for Research in Scientific Computation, & Center for Quantitative Studies in Biomedicine
North Carolina State University, Raleigh, NC 27695-8212

[2] Department of Mathematics & Statistics, Arizona State University
Tempe, AZ 85287-1804

[3] Mathematical, Computational and Modeling Sciences Center, Arizona State University
Tempe, AZ 85287-1904

[4] School of Human Evolution and Social Change, Arizona State University
Tempe, AZ 85287-2404

[5] Santa Fe Institute
Santa Fe, NM 87501

November 11, 2008

Abstract

The design and evaluation of epidemiological control strategies is central to public health policy. The collection and use of surveillance data is vital in this task, and is an area in which mathematical models and the use of inverse problem methods can have an impact. We apply these methods to the study of pneumococcal vaccination strategies as a relevant example which poses many challenges common to other infectious diseases. Methods for the calibration of an age-structured population model via parameter estimation are presented. We demonstrate that relevant yet typically unknown parameters may be estimated from infection and colonization prevalence data, which suggests that colonization information should be routinely collected. The calibrated model is used to assess implemented vaccine policies through the estimation of parameters if vaccine history is recorded along with infection and colonization information. Finally, we show how one might

*Corresponding author: Tel: 919.513.7084; Fax: 919.515.1636; E-mail address: klsutton@ncsu.edu

determine an appropriate level of refinement or aggregation in the age-structured model given age-stratified observations. These results illustrate ways in which the collection and analysis of surveillance data can be improved using inverse problem methods.

1 Introduction

Mathematical models have proven a beneficial tool for designing and evaluating prevention and treatment policies, particularly those associated with infectious diseases [2, 10]. Models are used to theoretically compare the effects of targeting prevention versus treatment, and some models have incorporated more sophisticated effects such as age-structure [15], when relevant. Inverse problem methods have been used with surveillance data to calibrate structured population models [5, 6, 8, 9]. These calibrated models can then be used to make specific predictions concerning the impact of intervention policies on a particular population. Inverse problem methods can also be used to assess the effectiveness of policies once in place, and to determine the level of complexity needed in the modeling framework as warranted by the data. We demonstrate the use of mathematical modeling and inverse problem methods in conjunction with surveillance data as useful tools that should be put systematically in the hands of public health officials.

Here we illustrate the power of these approaches in the context of pneumococcal diseases, or infections caused by *Streptococcus pneumoniae*. Although isolated in 1881 [3] the infections caused by these bacteria are still a significant cause of morbidity and mortality, with estimates of 1 million annual deaths from pneumococcal pneumonia occuring in children under the age of five [28]. The development of pneumococcal vaccines is an active research area, raising questions concerning the design of effective strategies implementing novel vaccines. While the recent licensing of the PCV7 vaccine has been encouraging in light of the drastic reduction seen in infections, the long-term effects of its widespread use are unclear. In fact, an immunization program in Australia in which all children were provided the PCV7 free of charge (and 90% coverage obtained), was shown in [25] to be decreasingly effective during the first three years by similar methods employed here. The dynamics of these infections are dependent on many factors - notably, age, nutrition, climate, and the prevalence of serotypes endemic to the geographic region. As such, it is not likely that one strategy will be effective for all populations. Thus the use of the approaches in this manuscript using pneumococcal surveillance data to design and assess such strategies is particularly important.

New infections are the most common reported data type, however, the asymptomatic nasopharyngeal colonization that precedes infection is the stage during which horizontal spread occurs. Colonization is usually reversed in healthy individuals and does not follow the same age-prevalence profile as is seen with infections [1, 12, 20]. Some vaccines in development have the potential to protect against colonization by some or all of the over 90 pneumococcal serotypes and the monitoring of colonization prevalence has been considered of increasing importance. Whether to target either colonization, infection, or a combination of both is unclear (see [26]). It is thought that targeting colonization may provide a previously unavailable ecological niche to other, potentially more invasive colonizers, not necessarily *S. pneumoniae*. Thus, programs using such vaccines would require constant monitoring, which can be effectively done using this framework.

In this paper, we first introduce a mathematical model of pneumococcal disease dynamics in which the population is structured by age in Section 2, and then describe typically known parameters along with sources in Section 3. The 'data' used here is generated from a forward solution of the mathematical model as described in Section 4.1. The inverse problem methods employed here are outlined in Section 4. We next describe the calibration of a mathematical model (Section 5) and the evaluation of an implemented control strategy (Section 6) that relies on parameter estimates from surveillance data. We investigate the uses of both infection and colonization data before and after the simulation of a novel vaccine policy. We discuss the parameters that can be estimated from each type of data. Further, we illustrate the interpretation of standard errors of parameters as a means of determining the number and frequency of longitudinal observations necessary to obtain reliable estimates.

We show that infection rates can be estimated from case notification data and that the reduction of cases by vaccination can be quantified if vaccination history in infected individuals is also available. Colonization prevalence data allows for the estimation of the force of infection, and if we are willing to assume proportionate mixing, the estimation of the age-specific effective contact matrix. If vaccination status of colonized individuals is also provided, we show that the reduction of colonization events due to vaccination can be estimated. These results are encouraging since the contact parameters, which govern the horizontal spread of infections through a population, are typically unmeasurable by other more direct approaches. Thus they are typically unavailable in scientific literature. Also, studies attempting to determine the effect of vaccination on the colonization prevalence of a population are controversial. The ability to reliably estimate this parame-

ter would be a significant advantage in the assessment of the effect of the current and developing vaccines. This parameter is of particular interest as it may indicate whether the vaccine is inducing undesirable evolutionary changes in the endemic pneumococci or competing colonizing species, potentially changing the clinical picture of these infections.

In practice, it is not clear which aspects of reported surveillance data, such as location, age, sex, weight, etc., would prove to be useful. The discussion in Section 7 is intended to illustrate tools which can be used to guide the appropriate level of sophistication in theoretical studies, and potentially in data collection. In some cases, ignoring age-dependent effects can result in a loss of information. In others, the inclusion of effects that are not warranted in the data can result in unnecessary sophistication and computational difficulty. We use simulated data, aggregated in non-uniform age classes to illustrate methodology for the determination of the optimal level of refinement necessary. In particular, we demonstrate the use of a model comparison statistic, described in Section 4.3 to help determine when the structure of the population should be explicitly incorporated and when it can be ignored.

2 Age-Structured Model of Pneumococcal Disease Dynamics

In this section, we formulate a model in which $n(a,t)$ (in units of "number per age") denotes the density of individuals of age a, $a \in [0,\infty)$. This model and the underlying biological assumptions are described in more detail in [26]. These individuals are classified by their infection state at age a and time t. Individuals are considered either susceptible $S(a,t)$ to pneumococcal infection, asymptomatically colonized $E(a,t)$ by *S. pneumoniae*, or infected $I(a,t)$. Both susceptible $S(a,t)$ and asymptomatically colonized individuals $E(a,t)$ are effectively vaccinated at the age-dependent per capita vaccination rate $\phi(a)$, that is, they move to vaccinated states $S_V(a,t)$ and $E_V(a,t)$, respectively, at a rate proportional to $\phi(a)$. Vaccination is not always completely protective, even against vaccine-included serotypes, thus there are also infected vaccinated individuals $I_V(a,t)$.

Transmission of *S. pneumoniae* occurs through respiratory droplets, and therefore the colonization process occurs as a result of an *effective contact* between a susceptible (or vaccinated susceptible) of age a at time t and any colonized or infected individual of age a' at time t, a process reflected in the effective contact rate $c(a,a')$.

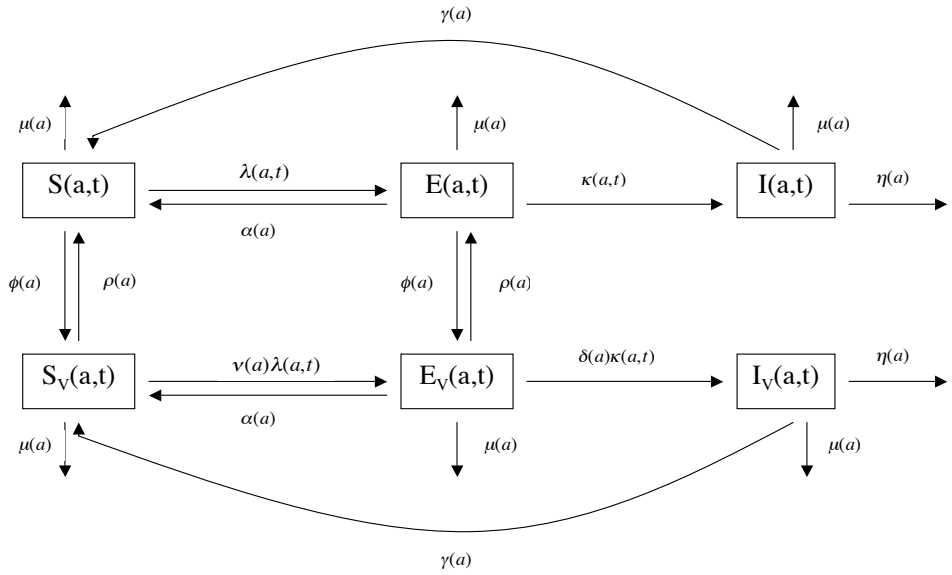

Figure 1: Pneumococcal infection dynamics with vaccination as a function of age.

The selected age-specific colonization rate $\lambda(a,t)$ used here is given by

$$\lambda(a,t) = \frac{\int_0^\infty c(a,a')\left[E(a',t) + E_V(a',t) + I(a',t) + I_V(a',t)\right] da'}{\int_0^\infty n(a',t) da'} \qquad (1)$$

where $\lambda(a,t)$ has units of $\frac{1}{\text{time}}$. Colonization is typically reversed at the age-specific per capita rate $\alpha(a)$, or individuals progress to infection at the seasonal per capita rate $\kappa(a,t)$. In vaccinated individuals colonization events are reduced by the factor $1-\nu(a)$ and infections are reduced by the factor $1-\delta(a)$, although the protection can be lost, represented by the per capita rate $\rho(a)$.

We assume that all individuals are born susceptible and unvaccinated, so the boundary conditions for the model are $E(0,t) = I(0,t) = S_V(0,t) = E_V(0,t) = I_V(0,t) = 0$ and $S(0,t) = \int_0^\infty f(a')n(a',t)da'$ where $f(a')$ is the age-specific per capita fertility rate.

The model equations are given by the following initial boundary value problem:

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)S(a,t) = -\lambda(a,t)S(a,t)+\alpha(a)E(a,t)+\gamma(a)I(a,t) \tag{2}$$
$$+\rho(a)S_V(a,t)-(\phi(a)+\mu(a))S(a,t),$$
$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)E(a,t) = \lambda(a,t)S(a,t)+\rho(a)E_V(a,t) \tag{3}$$
$$-(\alpha(a)+\kappa(a,t)+\phi(a)+\mu(a))E(a,t),$$
$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)S_V(a,t) = -\nu(a)\lambda(a,t)S_V(a,t)+\alpha(a)E_V(a,t) \tag{4}$$
$$+\gamma(a)I_V(a,t)+\phi(a)S(a,t)$$
$$-(\rho(a)+\mu(a))S_V(a,t),$$
$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)E_V(a,t) = \nu(a)\lambda(a,t)S_V(a,t)+\phi(a)E(a,t) \tag{5}$$
$$-(\alpha(a)+\delta(a)\kappa(a,t)+\rho(a)+\mu(a))E_V(a,t),$$
$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)I(a,t) = \kappa(a,t)E(a,t)-(\gamma(a)+\eta(a)+\mu(a))I(a,t), \tag{6}$$
$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)I_V(a,t) = \delta(a)\kappa(a,t)E_V(a,t), \tag{7}$$
$$-(\gamma(a)+\eta(a)+\mu(a))I_V(a,t),$$
$$S(0,t) = \int_0^\infty f(a')N(a',t)da',$$
$$E(0,t) = S_V(0,t) = E_V(0,t) = I(0,t) = I_V(0,t) = 0.$$

Model (2 – 7), while biologically reasonable, may be unnecessarily difficult framework if the goal is to explore the impact of age-specific vaccination. The Appendix of [26] contains a derivation of a system of ordinary differential equations (ODEs) to approximate system (2) – (7). The approach facilitates the aggregation of individuals into m age groups and is outlined in [15] and applied in [16] and [17]. This gives the system:

$$\frac{dS_1}{dt} = \sum_{j=1}^{m} f_j P_j - \lambda_1(t) S_1 + \alpha_1 E_1 + \gamma_1 I_1 + \rho_1 S_{V1} \tag{8}$$
$$- (\phi_1 + \mu_1 + b_1) S_1$$
$$\frac{dE_1}{dt} = \lambda_1(t) S_1 + \rho_1 E_{V1} - (\alpha_1 + \kappa_1(t) + \phi_1 + \mu_1 + b_1) E_1 \tag{9}$$
$$\frac{dS_{V1}}{dt} = -\nu_1 \lambda_1(t) S_{V1} + \alpha_1 E_{V1} + \gamma_1 I_{V1} + \phi_1 S_1 - (\rho_1 + \mu_1 + b_1) S_{V1} \tag{10}$$
$$\frac{dE_{V1}}{dt} = \nu_1 \lambda_1(t) S_{V1} + \phi_1 E_1 - (\alpha_1 + \delta_1 \kappa_1(t) + \rho_1 + \mu_1 + b_1) E_{V1} \tag{11}$$
$$\frac{dI_1}{dt} = \kappa_1(t) E_1 - (\gamma_1 + \eta_1 + \mu_1 + b_1) I_1 \tag{12}$$
$$\frac{dI_{V1}}{dt} = \delta_1 \kappa_1(t) E_{V1} - (\gamma_1 + \eta_1 + \mu_1 + b_1) I_{V1}, \tag{13}$$

$i = 2, .., m-1:$
$$\frac{dS_i}{dt} = -\lambda_i(t) S_i + \alpha_i E_i + \gamma_i I_i + \rho_i S_{Vi} \tag{14}$$
$$- (\phi_i + \mu_i + b_i) S_i + b_{i-1} S_{i-1} \tag{15}$$
$$\frac{dE_i}{dt} = \lambda_i(t) S_i + \rho_i E_{Vi} - (\alpha_i + \kappa_i(t) + \phi_i + \mu_i + b_i) E_i + b_{i-1} E_{i-1} \tag{16}$$
$$\frac{dS_{Vi}}{dt} = -\nu_i \lambda_i(t) S_{Vi} + \alpha_i E_{Vi} + \gamma_i I_{Vi} + \phi_i S_i \tag{17}$$
$$- (\rho_i + \mu_i + b_i) S_{Vi} + b_{i-1} S_{Vi-1}$$
$$\frac{dE_{Vi}}{dt} = \nu_i \lambda_i(t) S_{Vi} + \phi_i E_i \tag{18}$$
$$- (\alpha_i + \delta_i \kappa_i(t) + \rho_i + \mu_i + b_i) E_{Vi} + b_{i-1} E_{Vi-1}$$
$$\frac{dI_i}{dt} = \kappa_i(t) E_i - (\gamma_i + \eta_i + \mu_i + b_i) I_i + b_{i-1} I_{i-1} \tag{19}$$
$$\frac{dI_{Vi}}{dt} = \delta_i \kappa_i(t) E_{Vi} - (\gamma_i + \eta_i + \mu_i + b_i) I_{Vi} + b_{i-1} I_{Vi-1}, \tag{20}$$

$i = m$:

$$\frac{dS_m}{dt} = -\lambda_m(t)S_m + \alpha_m E_m + \gamma_m I_m + \rho_m S_{Vm} \tag{21}$$

$$- (\phi_m + \mu_m)S_m + b_{m-1}S_{m-1} \tag{22}$$

$$\frac{dE_m}{dt} = \lambda_m(t)S_m + \rho_m E_{Vm} \tag{23}$$

$$- (\alpha_m + \kappa_m(t) + \phi_m + \mu_m)E_m + b_{m-1}E_{m-1}$$

$$\frac{dS_{Vm}}{dt} = -\nu_m\lambda_m(t)S_{Vm} + \alpha_m E_{Vm} + \gamma_m I_{Vm} + \phi_m S_m \tag{24}$$

$$- (\rho_m + \mu_m)S_{Vm} + b_{m-1}S_{Vm-1}$$

$$\frac{dE_{Vm}}{dt} = \nu_m\lambda_m(t)S_{Vm} + \phi_m E_m \tag{25}$$

$$- (\alpha_m + \delta_m\kappa_m(t) + \rho_m + \mu_m)E_{Vm} + b_{m-1}E_{Vm-1}$$

$$\frac{dI_m}{dt} = \kappa_m(t)E_m - (\gamma_m + \eta_m + \mu_m)I_m + b_{m-1}I_{m-1} \tag{26}$$

$$\frac{dI_{Vm}}{dt} = \delta_m\kappa_m(t)E_{Vm} - (\gamma_m + \eta_m + \mu_m)I_{Vm} + b_{m-1}I_{Vm-1}. \tag{27}$$

All age-dependent rates ($r(a)$) are approximated by piecewise constant functions with at most $m-1$ discontinuities,

$$r(a) \approx \sum_{i=1}^{m} r_i \chi_{[a_{i-1}-a_i]},$$

where χ_i is the characteristic function on the age interval $[a_{i-1}, a_i]$. The length of the age intervals can vary. Smaller age intervals are taken in ranges where the dynamics of infections are likely changing drastically, i.e., the younger and older age ranges in the case of pneumococcal diseases.

3 'Known' or typically available parameters

There are a number of parameter values available in scientific literature and demographic or census reports, which have been discretized for age where appropriate. These values, along with sources, are summarized in Table 1. If the value of a particular parameter is likely population-specific, as is the case for mortality and birth rates, we chose rates that result in a stable age distribution. We note that these rates are typically available

in demographic reports for most populations. We consider five age groups: $[0, 2], (2, 15], (15, 50], (50, 65]$, and $(65, \infty)$ for the parameter estimation studies.

Although there are no reliable reported values for the recovery rates, γ_i, our previous work with the unstructured model [25], suggests that values within the physiologically feasible range $\gamma_i \in [\frac{1}{2}, 2]$, do not change results for the other estimated parameters. Due to lack of information we are not motivated to consider age-dependence of this parameter and therefore set $\gamma = \gamma_1 = ... = \gamma_5$. We show age-specific parameter estimates are also not affected for γ fixed at values on the boundaries of the interval $[\frac{1}{2}, 2]$.

When simulating a population before the implementation of a new vaccine policy, portions of age groups 4 and 5 are vaccinated with the polysaccharide vaccine (PPV23), in accordance with the vaccination policies of most developed countries since the mid-1980s. Since this vaccine has had no observed effect on carriage, we set $\nu_i = 1$ for $i = 1, ..., 5$ in the model calibration studies. Parameters of interest to be estimated are: mean infection rates $\kappa_0 = \{\kappa_{0,i}\}_{i=1}^5$, vaccine protection from infection δ_4, δ_5, and the force of infection $\lambda(t) = \{\lambda_i(t)\}_{i=1}^5$. These are parameters that significantly govern the epidemiological dynamics of pneumococcal diseases, yet are not available in most scientific sources. We have chosen values for these parameters that produce annual infections by age that correlate well with those available on the Australian NNDS website [12]. Additionally, these rates result in reasonable values for age-specific prevalences, based on scientific reviews and primary research papers, such as [1, 11, 19, 27]. The infections and colonization prevalences for these age groups as compared to the reported values are shown in Figure 2. The initial conditions were chosen such that neither the age distribution nor the colonization prevalence changes significantly with time. These parameter values, which will be taken as the 'true' values for the estimation studies, along with the initial conditions are listed in Table 2.

To simulate a newly implemented vaccine policy, targeting children, we increase the vaccination rate of age group 1. The pneumococcal conjugate vaccine is given at the ages of 2, 4, and 6 months, with a dose of PPV23 sometimes recommended for children aged 18-24 months. Since the booster is done rarely and therefore does not likely affect the infection dynamics at the population level, it is not specifically incorporated in these studies. In practice, one should use age intervals in the model to reflect the vaccine program of interest. We have not done that here since the purpose of this work is largely to illustrate the methodology and no *specific* vaccine policy is considered. Values for ρ have been chosen to consider vaccination protection

Table 1: Available parameters for discrete age-structured model with 5 age classes. General information is shown for age-specific parameters, $r(a)$, and specific values r_i where $i = 1, .., 5$ for a 5-age class model are given. If no value is shown, $r_i = 0$. Sources denoted by a * are places where parameter values may typically be found. The values shown for these parameters have not been fixed according to any particular source.

Parameter	Source/ Value
$f(a)$	Demographic/census reports*
f_3	0.01825
$\mu(a)$	Demographic/census reports*
μ_1	$3.889e^{-4}$
μ_2	$1.112e^{-5}$
μ_3	$1.3e^{-5}$
μ_4	$4e^{-4}$
μ_5	0.0075
γ	Not available
γ	1
$\alpha(a)$	[11]
α_1	0.25
α_2	1
α_3	1.5
α_4	1.5
α_5	1.5
$\eta(a)$	[21, 22, 23]
η_1	$8.33e^{-4}$
η_2	0.00417
η_3	0.00583
η_4	0.0075
η_5	0.015

Table 2: Parameters and initial conditions used to generate data for model calibration studies.

Parameter	Value	Parameter	Value
$\kappa_{0,1}$	$3e^{-4}$	δ_1	1
$\kappa_{0,2}$	$2.5e^{-5}$	δ_2	1
$\kappa_{0,3}$	$4e^{-5}$	δ_3	1
$\kappa_{0,4}$	$6e^{-5}$	δ_4	0.8
$\kappa_{0,5}$	$1.7e^{-4}$	δ_5	0.6
ϕ_4	0.001	ρ_4	0.01583
ϕ_5	0.03	ρ_5	0.015
$S_1(t_0)$	371,915	$S_2(t_0)$	2,826,193
$E_1(t_0)$	190,489	$E_2(t_0)$	818751
$S_{V1}(t_0)$	0	$S_{V2}(t_0)$	0
$E_{V1}(t_0)$	0	$E_{V2}(t_0)$	0
$I_1(t_0)$	32	$I_2(t_0)$	13
$I_{V1}(t_0)$	0	$I_{V2}(t_0)$	0
$S_3(t_0)$	8,522,665	$S_4(t_0)$	3,271,002
$E_3(t_0)$	1,225,498	$E_4(t_0)$	454,933
$S_{V3}(t_0)$	0	$S_{V4}(t_0)$	150,127
$E_{V3}(t_0)$	0	$E_{V4}(t_0)$	20,877
$I_3(t_0)$	28	$I_4(t_0)$	16
$I_{V3}(t_0)$	0	$I_{V4}(t_0)$	1
$S_5(t_0)$	1,043,021		
$E_5(t_0)$	176,005		
$S_{V5}(t_0)$	1,426,475		
$E_{V5}(t_0)$	241,114		
$I_5(t_0)$	17		
$I_{V5}(t_0)$	14		

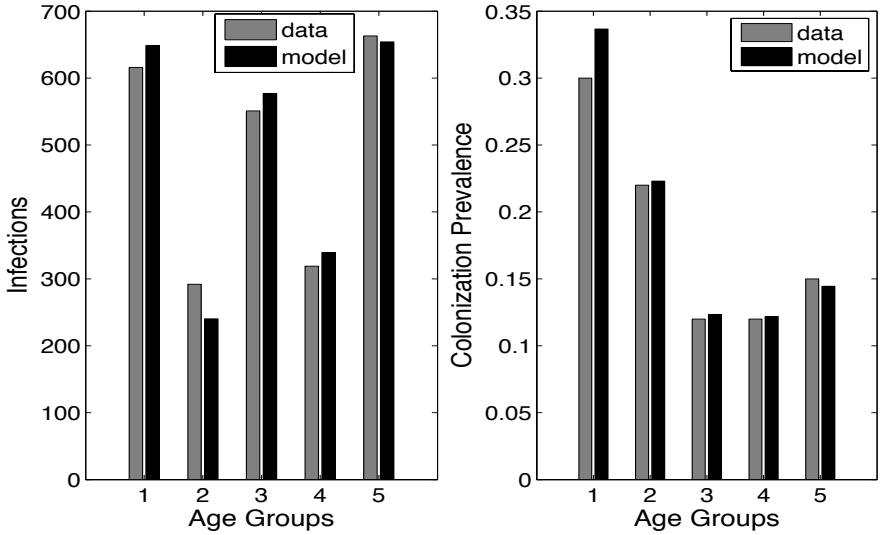

Figure 2: Reported and model calculated infections (left) and colonization prevalences (right) by age.

only and excludes the natural immunity that typically develops as a result of exposure to pneumococcus over time. We update the values of δ_1, δ_2 and ν_1, ν_2 to reflect a vaccine which has a protective effect against infection in the first two age groups, and a noticeable impact on colonization in the first three age groups. This effect has been observed in some populations in which the PCV7 has been implemented, and although it is still controversial, other vaccines in development have the potential to have a stronger protective effect against the colonization process. Initial conditions are the same as those used for the model calibration studies, since both colonization and infections are endemic. Therefore, it is reasonable to assume that this is a realistic starting point for both the model calibration and the population before the implementation of a novel vaccination program. The initial conditions and the infection rates and effective contact rates are unchanged from the model calibration studies and are therefore excluded from Table 3, which contains the 'true' parameter values that have been used to generate data corresponding to the implementation of a childhood vaccine.

Table 3: Parameters used to generate data simulating the implementation of a novel vaccine program.

Parameter	Value	Parameter	Value
δ_1	0.2	ν_1	0.3
δ_2	0.4	ν_2	0.5
δ_3	1	ν_3	1
δ_4	0.8	ν_4	1
δ_5	0.6	ν_5	1
ϕ_1	0.9	ρ_1	0
ϕ_2	0.05	ρ_2	$\frac{1}{24}$
ϕ_3	0	ρ_3	0
ϕ_4	0.001	ρ_4	0.01583
ϕ_5	0.03	ρ_5	0.015

4 Inverse problem methods

4.1 Surveillance data description

We generate data to simulate the kind of information that is available or measurable by surveillance programs. These data are used in conjunction with the mathematical model to estimate parameters via inverse problem methodology. Some of the data generated are collected more commonly than others, and there have been recent discussions on which information should be reported, therefore it is valuable to explore the uses of each type of information. Total cases are commonly reported, although the age of the individual or vaccination status is not always recorded. Vaccine efficacy studies occasionally incorporate nasal swabs of the studied population, which reveals the proportion of individuals colonized by *S. pneumoniae*. Recently, it has been suggested that public health departments should collect this kind of information on a regular basis and at a large-scale to determine if any protective or potentially harmful effects have occurred. We discuss some benefits of these additional efforts as they relate to the process of assessing the impact of a vaccine policy on a surveyed population.

We generate data according to the following forms:

- Total cases: $f_j^{(1,i)} = \int_{t_j}^{t_{j+1}} \left[\kappa_i(s)E_i(s) + \delta_i\kappa_i(s)E_i(s)\right] ds$ where $j = 1, ..., n_1$ for each $i = 1, ..., 5$,

- Vaccinated Cases: $f_j^{(2,i)} = \int_{t_j}^{t_{j+1}} \delta_i \kappa_i(s) E_{Vi}(s) ds$ where $j = 1, .., n_2$ for each $i = 1, ..., 5$,
- Colonization Prevalence: $f_j^{(3,i)} = \frac{E_i(t_j) + E_{Vi}(t_j)}{N_i}$ where $j = 1, ..., n_3$ for each $i = 1, ..., 5$,
- Vaccinated Colonized Prevalence: $f_j^{(4,i)} = \frac{E_{Vi}(t_j)}{N_i}$ where $j = 1, ..., n_4$ for each $i = 1, .., 5$.

Since surveillance data typically contains some observational error, noise was added to the generated data according to the statistical model

$$Y_j^{(k,i)} \sim f^{(k,i)}(t_j, \theta_0) + \epsilon_j^{(k,i)} \tag{28}$$

where $Y_j^{(k,i)}$ represents data of type k at time t_j for age class i. Here $f^{(k,i)}(t_j, \theta_0)$ is the model quantity corresponding to the data with parameters θ_0, and $\epsilon_j^{(k,i)}$ is a random variable representing observational error. Note that since $Y_j^{(k,i)}$ depends on $\epsilon_j^{(k,i)}$, it is also a random variable, and one set of data is a *realization* $y_j^{(k,i)}$ of the random variable $Y_j^{(k,i)}$. In this case we know the true parameters θ_0, but in general we do not and instead we assume that such a parameter exists. That is, we assume that the mathematical model is essentially correct, and any difference from the model prediction and the observed data is due to observational error, as modeled by $\epsilon_j^{(k,i)}$. The error is sampled from a normal distribution, $\epsilon_j^{(k,i)} \sim \sigma_{k,i} * \mathcal{N}(0,1) = \mathcal{N}(0, \sigma_{k,i}^2)$. In practice, we do not typically know the distribution of the error, and the least squares approach taken here does not assume one. However, one can usually make some reasonable assumptions concerning the first two moments of the error (commonly that $E[\epsilon_j] = 0$ and $var[\epsilon_j] = \sigma_0^2$ for $j = 1, ..., n$), and this is enough to proceed. The variance of the error $var[\epsilon_j^{(k,i)}] = \sigma_{k,i}^2$ is scaled by the magnitude of the observations, but is constant in time, so $\epsilon_j^{(k,i)}$ are $i.i.d.$ for $j = 1, ..., n_k$, where n_k is the number of longitudinal observations of type k. For further discussion on estimation procedures and the consequences of assumptions on the random variable $\epsilon_j^{(k,i)}$ observational error, see [4]. The values of the variances for the purposes of generating data are determined by

$$\sigma_{k,i} = \frac{l}{100} * \text{average}_j(f_j^{(k,i)}) \tag{29}$$

where l is the 'level' of noise, or in other words, the above formula would give data with l% noise.

4.2 Ordinary least squares estimation

We discuss the estimation of parameters θ by an ordinary least squares (OLS) approach in which we minimize the difference between generated data and the model. In Section 5, the parameters estimated are $\theta = (\{\kappa_{0,i}\}_{i=1}^5, \{\delta_i\}_{i=4}^5, \{\lambda_i\}_{i=1}^5)$ and in Section 6 we estimate $\theta = (\{\nu_i\}_{i=1}^2, \{\delta_i\}_{i=1}^5)$. The OLS estimator $\theta_{OLS}(Y^{(k,i)})$ is given by

$$\theta_{OLS}(Y^{(k,i)}) = \arg\min_{\theta\in\Theta} \sum_{i=1}^{m} \frac{1}{\sigma_{k,i}^2} \sum_{j=1}^{n_k} \left[Y_j^{(k,i)} - f^{(k,i)}(t_j, \theta) \right]^2 \tag{30}$$

where Θ is a feasible parameter space contained in $\mathbb{R}^p$ and p is the number of parameters to be estimated. The estimator is a random variable depending on the random variable $Y_j^{(k,i)}$ and we seek to obtain estimates $\hat{\theta}$ using a realization or observed data $y_j^{(k,i)}$ for $j = 1, ..., n_k$.

We explore which types of data contain information on the above parameters. To illustrate the implementation of the OLS procedure, we outline here the estimation of $\{\kappa_{0,i}\}_{i=1}^5$ from 6 years of annual case notifications ($y^{(1)} = \{y^{(1,i)}\}_{i=1}^5$), a realization of $Y^{(1)} = \{Y^{(1,i)}\}_{i=1}^5$. The estimates $\hat{\theta} = (\{\kappa_{0,i}\}_{i=1}^5)$ of the true parameters θ_0 minimize the objective functional

$$J_{n_k}(\theta, y^{(1)}) = \sum_{i=1}^{5} \frac{1}{\sigma_{1,i}^2} \sum_{j=1}^{n_k} \left[y_j^{(1,i)} - f^{(1,i)}(t_j, \theta) \right]^2, \tag{31}$$

where the feasible parameter space is contained in $\theta \in \Theta \subset \mathbb{R}^5$ where $t_j \in \{0, 12, 24, 36, 48, 60, 72\}$ and $n_k = 6$. The variances in the observation error are given by

$$\hat{\sigma}_{1,i}^2 = \frac{1}{n_k - p} \sum_{j=1}^{n_k} \left[Y_j^{(1,i)} - f^{(1,i)}(t_j, \hat{\theta}) \right]^2. \tag{32}$$

With a set of data $y_j^{(k,i)}$ we estimate σ_0 by $\hat{\sigma}$ using Equation (32) with the estimates $\hat{\theta}$. Inspection of objective functional (31) reveals that $\hat{\theta}$ depends on the unknowns, $\sigma_1 = \{\sigma_{1,i}^2\}_{i=1}^5$, so we use the following iterative least squares estimation process to estimate $\psi = (\theta, \{\sigma_{1,i}^2\}_{i=1}^5)$:

1. Guess initial values for $\hat{\sigma}_{1,i}^2$ and solve for an initial estimate $\hat{\theta}^{(0)}$ using (31). Set $k = 1$.

2. Using $\hat{\theta} = \hat{\theta}^{(k)}$, calculate $\hat{\sigma}_{1,i}^2$ using the expression in (32).

3. Increment k by 1. Estimate $\hat{\theta}^{(k)}$ by minimizing (31) with $\hat{\sigma}^2_{1,i}$ from step 2.

4. Re-estimate $\hat{\sigma}^2_{1,i}$ by (32) with $\hat{\theta} = \hat{\theta}^{(k)}$.

5. Repeat steps 3 and 4, until two successive estimates for $\hat{\theta}$ are sufficiently close. Then take $\hat{\theta} = \hat{\theta}^{(k)}$ and the current estimate for $\hat{\sigma}^2_{1,i}$.

In the equation for $\hat{\sigma}_{k,i}$, the factor $\frac{1}{n_k - p}$ requires that the number of longitudinal data points for each type of observation for each age group be greater than the parameters to be estimated. Note that this implies that when estimating one age-dependent parameter discretized into m rates we require at least $m + 1$ longitudinal observations.

Estimates of the same parameters $\hat{\psi} = (\{\kappa_{0,i}\}_{i=1}^5, \{\sigma_{1,i}\}_{i=1}^5)$ using two years of monthly case notifications minimize the objective functional

$$J_{24}(\theta, y^{(1)}) = \sum_{i=1}^{5} \frac{1}{\sigma^2_{1,i}} \sum_{j=1}^{24} \left[y_j^{(1,i)} - f^{(1,i)}(t_j, \theta) \right]^2, \tag{33}$$

where now $t_j \in \{0, 1, 2, ..., 24\}$, and the expressions for the estimated variances in the observation error are

$$\hat{\sigma}^2_{1,i} = \frac{1}{24 - 5} \sum_{j=1}^{24} \left[y_j^{(1,i)} - f^{(1,i)}(t_j, \hat{\theta}) \right]^2. \tag{34}$$

Including additional data types can provide information allowing for the estimation of additional parameters in some cases. For example, estimating $\theta = (\{\kappa_{0,i}\}_{i=1}^5, \{\delta_i\}_{i=4}^5)$ via OLS with two years of total ($Y^{(1)}$) and vaccinated ($Y^{(2)}$) cases reported monthly involves minimizing

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{k=1,2} \sum_{i=1}^{5} \frac{1}{\sigma^2_{k,i}} \sum_{j=1}^{24} \left[y_j^{(k,i)} - f^{(k,i)}(t_j, \theta) \right]^2. \tag{35}$$

The expressions for $\hat{\sigma}^2_{1,i}$ and $\hat{\sigma}^2_{2,i}$ are the analogues of those shown in (32) and (34) and are given by

$$\hat{\sigma}_{1,i} = \frac{1}{24 - 7} \sum_{j=1}^{24} \left[y_j^{(1,i)} - f^{(1,i)}(t_j, \hat{\theta}) \right]^2,$$

$$\hat{\sigma}_{2,i} = \frac{1}{24 - 7} \sum_{j=1}^{24} \left[y_j^{(2,i)} - f^{(2,i)}(t_j, \hat{\theta}) \right]^2.$$

Once we have estimated $\hat{\psi} = (\hat{\theta}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$, we calculate standard errors as a measure of reliability of the estimates obtained. Under reasonable assumptions on smoothness and regularity (smoothness requirements are easily verified using continuous dependence results for differential equations in most cases, and regularity requirements involve, among others, conditions on the manner in which the sample size increases), the standard nonlinear regression approximation theory ([13], [14], [18], and Ch 12 of [24]) for asymptotic (as $n \to \infty$) distributions can be invoked. By this theory, the sampling distribution for θ_{OLS} is approximately a p-multivariate Gaussian with mean $E[\theta_{OLS}(Y)] = \theta_0$ and variance $var[\theta_{OLS}(Y)] \approx \Sigma_0 = \sigma_0^2[n\Omega_0]^{-1}$ where Ω_0 is defined as

$$\Omega_0 \equiv \lim_{n\to\infty} \frac{1}{n} \chi^T(\theta_0)\chi(\theta_0) \tag{36}$$

Under conditions outlined further in [24] we assume Ω_0 exists. Here $\chi(\theta)$ is the $n \times p$ sensitivity matrix with elements

$$\chi_{jl}(\theta) = \frac{\partial f(t_j, \theta)}{\partial \theta_l}. \tag{37}$$

That is, for n large, the sampling distribution is approximately,

$$\theta_{OLS}(Y) \sim \mathcal{N}_p(\theta_0, \Sigma_0) \approx \mathcal{N}_p(\theta_0, \sigma_0^2 \left[\chi^T(\theta_0)\chi(\theta_0)\right]^{-1}). \tag{38}$$

Then for the estimates $\hat{\theta}$, the covariance matrix Σ_0 is estimated using $\hat{\theta}, \hat{\sigma}^2$

$$\Sigma_0 \approx \hat{\Sigma} = \hat{\sigma}^2 \left[\chi^T(\hat{\theta})\chi(\hat{\theta})\right]^{-1}.$$

The vectorized analogue, accounting for q data types, of the above equation is

$$\hat{\Sigma} = \left(\sum_{j=1}^{n} D_j^T(\hat{\theta}) \hat{V}^{-1} D_j(\hat{\theta}) \right)^{-1}, \tag{39}$$

where $D_j(\theta)$ is a $q \times p$ matrix with the $(k*i, l)$th entry defined by $\frac{\partial f^{(k,i)}(t_j,\theta)}{\partial \theta_l}$ where k indexes the data type, i the age class, and l refers to the parameter, or element of θ. The matrix $\hat{V}$ is diagonal with entries of the variances of the observational errors $\hat{V} = \text{diag}(\hat{\sigma}_1^2, ..., \hat{\sigma}_k^2)$. The matrices Σ_0 and $\hat{\Sigma}$ are $p \times p$ square matrices and the standard error of parameter θ_l

$$SE(\hat{\theta}_l) \approx \sqrt{\hat{\Sigma}_{ll}}. \tag{40}$$

The standard errors are an indication of the *reliability* of the estimation procedure, but do not contain any information on the *correctness* of the estimates. That is, the standard errors should not be interpreted as a measure of 'closeness' to the true values θ_0 which are usually unknown. However, small standard errors are one indication that the data used in the estimation procedure contains a suitable amount of information on the parameters estimated and therefore provides some confidence in the estimates obtained.

For the example where total and vaccinated cases $(\{y^{(1,i)}\}_{i=1}^5, \{y^{(2,i)}\}_{i=1}^5)$ are used to estimate θ, $D_j(\theta)$ is a 10×7 matrix defined by

$$D_j(\theta) = \begin{pmatrix} \frac{\partial f^{(1,1)}(t_j,\theta)}{\partial \kappa_{0,1}} & \frac{\partial f^{(1,1)}(t_j,\theta)}{\partial \kappa_{0,2}} & \cdots & \frac{\partial f^{(1,1)}(t_j,\theta)}{\partial \delta_5} \\ \vdots & & & \vdots \\ \frac{\partial f^{(1,5)}(t_j,\theta)}{\partial \kappa_{0,1}} & \frac{\partial f^{(1,5)}(t_j,\theta)}{\partial \kappa_{0,2}} & \cdots & \frac{\partial f^{(1,5)}(t_j,\theta)}{\partial \delta_5} \\ \frac{\partial f^{(2,1)}(t_j,\theta)}{\partial \kappa_{0,1}} & \frac{\partial f^{(2,1)}(t_j,\theta)}{\partial \kappa_{0,2}} & \cdots & \frac{\partial f^{(2,1)}(t_j,\theta)}{\partial \delta_5} \\ \vdots & & & \vdots \\ \frac{\partial f^{(2,5)}(t_j,\theta)}{\partial \kappa_{0,1}} & \cdots & \cdots & \frac{\partial f^{(2,5)}(t_j,\theta)}{\partial \delta_5} \end{pmatrix}. \tag{41}$$

Then $\hat{V} = diag(\hat{\sigma}^2_{1,1}, \hat{\sigma}^2_{1,2}, ..., \hat{\sigma}^2_{1,5}, \hat{\sigma}^2_{2,1}, ..., \hat{\sigma}^2_{2,5})$. For a more general and detailed discussion of these methods see [4].

4.3 Model comparison statistic

Increasing the number of parameters from a given set of data results in a residual (value of the minimized objective functional) that is either the same or smaller than when fewer parameters are estimated. This is sometimes due to an improved fit of the model to the data and sometimes simply due to an increase in the degrees of freedom. In the case of one model being a special case of another, or a 'nested' model, there is a statistic which can be used to determine whether the reduced residual is statistically due to an improved fit.

Consider the estimator $\theta(Y)$ resulting from the minimization of model quantities $f(t_j, \theta)$ and observations Y_j for $j = 1, ..., n$:

$$\theta_{OLS}(Y) = \arg\min_{\theta \in \Theta} J_n(Y, \theta) \tag{42}$$

where $\Theta \subset \mathbb{R}^p$. We assume the same statistical model as before, $Y_j = f(t_j, \theta_0) + \epsilon_j$ for $j = 1, ..., n$. Again, we assume ϵ_j are *i.i.d.* for $j = 1, ..., n$ with mean $E[\epsilon_j] = 0$ and variance $var[\epsilon_j] = \sigma_0^2$. The quantity $\theta_0 \in \Theta$ again

represents the 'true' parameters which generated the observations. Let $\hat{\theta}$ denote the estimate obtained from a realization $y = \{y_j\}_{j=1}^n$ of Y in which the objective functional has been minimized over Θ.

It may be that the true parameter is just as likely to lie within a subset $\Theta_H \subset \Theta$ which corresponds to the parameter θ being subject to some constraints. The set Θ_H is defined by

$$\Theta_H = \{\theta \in \Theta | H\theta = c\} \tag{43}$$

where H is an $r \times p$ matrix of full rank, and c is a known constant. In the context of this paper, a model incorporating fewer age classes, and therefore having fewer dimensions of each age-dependent parameter, will result in a minimization over a smaller parameter space. Let θ^H denote the estimator given by

$$\theta^H(Y) = \arg\min_{\theta \in \Theta_H} J_n(Y, \theta) \tag{44}$$

where a realization is used to obtain $\hat{\theta}^H$. We would like to test the null hypothesis

$$H_0 : \theta_0 \in \Theta_H. \tag{45}$$

The test statistic $U_n^r(Y)$ is defined by

$$U_n^r(Y) = \frac{n(J_n(Y, \theta(Y)) - J_n(Y, \theta^H(Y)))}{J_n(Y, \theta(Y))}. \tag{46}$$

Asymptotic convergence results for $U_n(Y)$ similar to those in asymptotic sampling theory [7, 8] are established in [24]. Then if H_0 is true, U_n converges in distribution to $U(r)$ as $n \to \infty$ where $U(r) \sim \chi^2(r)$, a χ^2 distribution with r degrees of freedom.

Then for a given significance level α, $Prob\{U_n^r > \tau\} = \alpha$ for a threshold τ. We reject H_0 as false if $\hat{U}_n^r > \tau$, or we do not reject H_0 as true if $\hat{U}_n^r < \tau$ at significance level α.

For example, consider the estimation of parameters from a model considering 8 age classes, and thus $\hat{\theta}^{(8)} = (\kappa_{0,1}, ..., \kappa_{0,8})^T$, compared with parameter estimates $\hat{\theta}^{(6)} = (\kappa_{0,1}, ..., \kappa_{0,6})^T$ from a model considering 6 age classes such that the first three age classes in the more discretized model are considered equivalent. That is, $\hat{\theta}^{(6)}$ is a special case of $\hat{\theta}^{(8)}$ with $\kappa_{0,1} = \kappa_{0,2} = \kappa_{0,3}$. Then the 2 x 8 matrix H is given by

$$H = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and $c = (0,0)^T$. The test statistic

$$\hat{U}_n^2(y) = \frac{n(J_n(y,\hat{\theta}^6) - J_n(y,\hat{\theta}^8))}{J_n(y,\hat{\theta}^8)}$$

is then compared to a $\chi^2(2)$ distribution. For a significance level α, we reject $H_0 : \theta_0 \in \Theta_H \subset \mathbb{R}^6$ if $\hat{U}_n > \tau$. That is, the age-dependent effects in the first three age classes do appear to provide a significant improvement in the fit to data with a confidence level of $(1-\alpha) * 100\%$.

5 Model calibration

Here we discuss the estimation of relevant model parameters from collectable surveillance data to calibrate a model to a particular population. We focus on rates that are not typically available in the literature, such as the mean infection rates $\kappa_{0,i}, i = 1, ..., 5$, or in some cases, rates that are not directly measurable. Specifically, we focus on a population in which the polysaccharide vaccine is currently implemented, and thus the effect of this vaccine on the older age groups (4 and 5) is incorporated, and estimates of δ_4, δ_5 discussed. There is not enough information in the data considered here to estimate the age-specific contact rates $c_{i,j}, i, j \in \{1, ..., 5\}$, but since the colonization prevalence does not change with time, or in other words, this state is endemic, we consider the force of infection as being constant over a short time. So taking $\Lambda_i(t) \approx \Lambda_i$ for each $i = 1, ..., 5$, we are able to estimate the age-specific forces of infection in a population in which the vaccine implemented does not protect against colonization. The recovery rate γ is also not estimated from the data considered, but we see that estimating other pertinent parameters when the recovery rates are fixed at values $\gamma = \frac{1}{2}, \gamma = 1$, or $\gamma = 2$ does not affect the other estimates.

From reports of the total number of cases of invasive pneumococcal diseases, we are able to estimate the mean infection rates, $\kappa_0 = \{\kappa_{0,i}\}_{i=1}^5$. Table 4 contains the results of estimating these rates from data collected annually for 6 years with 0 and 10% noise (generated according to the statistical model (28)). At least six time points are required for the estimation of these parameters, due to the formulas for $\sigma_{k,i}$ (the formula for each $\sigma_{k,i}$ contains a factor $\frac{1}{n_k - p}$ where p is the number of parameters to be estimated). It may not be realistic to assume that surveillance data has been collected

over a long period of time. But as we can see by the parameter values and standard errors in the last two columns of Table 4, using more frequently reported data (monthly) over a shorter period of time produces comparable results. In fact, the parameter estimates are improved, as their corresponding standard errors are reduced. The model solution fitted to the monthly generated data is shown in Figure 3.

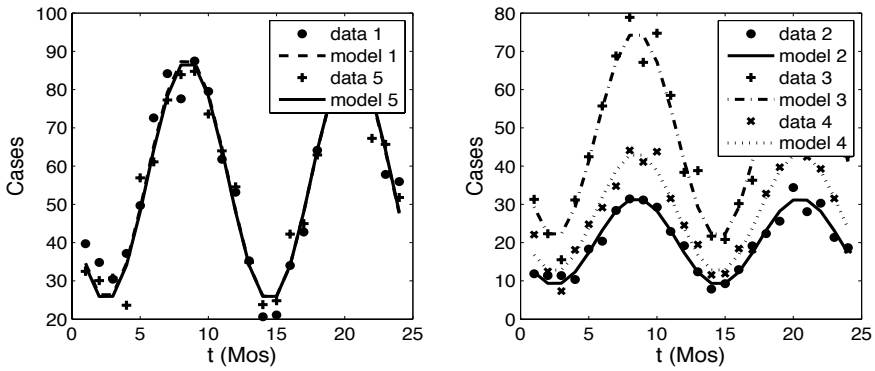

Figure 3: Model fit to 2 years of monthly case notifications with 10% noise.

Using the number of vaccinated cases $Y^{(2)}$ in addition to the number of total cases $Y^{(1)}$ reported by month, we are able to estimate the factors δ_4, δ_5 representing the vaccine-induced protection from infection, in addition to the mean infection parameters, as shown in Table 5 and Figure 4. Annual data are not used in this case since we are estimating seven parameters simultaneously and would therefore need at least eight years of data, which is likely not available in most cases. We are also able to estimate these rates when vaccinated ($Y^{(2)}$) and unvaccinated cases ($Y^{(1)} - Y^{(2)}$) are used in the OLS procedure with similar results. That is, the parameter values are comparably close to their true values, and the corresponding standard errors are of similar magnitude. Since there are no notable differences from the estimation, the results are not shown.

Effective contact rates are difficult, if not impossible, to measure directly, and therefore are not typically available in literature. Since collectively, these rates largely govern the level of colonization observed in a population, the colonization prevalence data $Y^{(4)}, Y^{(5)}, Y^{(6)}$ is most likely to contain the most information on the contact rates. However, when these data are used in the estimation process, the sensitivity of the data to the parameters is approximately zero, resulting in standard errors too large to be calculated due to a numerically singular matrix $\chi^T\chi$. It is reasonable to approximate

Table 4: Estimates of $\psi = (\{\kappa_{0,i}\}_{i=1}^5, \{\sigma_{1,i}\}_{i=1}^5)$ from 6 years of annual case notification data ($Y^{(1)}$) with 0 and 10% noise (where the level of noise is denoted by superscripts), are shown in the top portion of the table in columns 3-6. Estimates of ψ from 2 years of monthly case notification data with 10% noise is shown in the bottom portion of the table in columns 5 and 6. The true values, which were used to generate the data, are denoted by ψ_0, and are shown in column 2. Note that the values for $\sigma_{1,i}^{10}$ for the annual data are significantly larger than those for the monthly data since they are scaled by the observations and there are significantly more infections that occur in a year than a month.

	ψ_0	Annual, 6 yrs $\hat{\psi}^0$	$SE(\hat{\theta}^0)$	Annual, 6 yrs $\hat{\psi}^{10}$	$SE(\hat{\theta}^{10})$
$\kappa_{0,1}$	$3e^{-4}$	$3e^{-4}$	$6.0e^{-20}$	$3.22e^{-4}$	$4.9e^{-6}$
$\kappa_{0,2}$	$2.5e^{-5}$	$2.5e^{-5}$	$1.5e^{-20}$	$2.56e^{-4}$	$1.1e^{-6}$
$\kappa_{0,3}$	$4e^{-5}$	$4e^{-5}$	$3.7e^{-20}$	$3.89e^{-4}$	$2.7e^{-6}$
$\kappa_{0,4}$	$6e^{-5}$	$6e^{-5}$	$1.0e^{-19}$	$5.79e^{-5}$	$5.6e^{-6}$
$\kappa_{0,5}$	$1.7e^{-4}$	$1.7e^{-4}$	$3.5e^{-19}$	$1.82e^{-4}$	$1.9e^{-5}$
$\sigma_{1,1}^0$	0	$1.8e^{-12}$		N/A	
$\sigma_{1,2}^0$	0	$6.7e^{-13}$		N/A	
$\sigma_{1,3}^0$	0	$1.3e^{-12}$		N/A	
$\sigma_{1,4}^0$	0	$1.0e^{-12}$		N/A	
$\sigma_{1,5}^0$	0	$2.3e^{-12}$		N/A	
$\sigma_{1,1}^{10}$	149	N/A		148.6287	
$\sigma_{1,2}^{10}$	51	N/A		51.2244	
$\sigma_{1,3}^{10}$	96	N/A		96.2737	
$\sigma_{1,4}^{10}$	54	N/A		54.4548	
$\sigma_{1,5}^{10}$	127	N/A		127.2959	
				Monthly, 2 yrs $\hat{\psi}^{10}$	$SE(\hat{\theta}^{10})$
$\kappa_{0,1}$	$3e^{-4}$			$3.00e^{-4}$	$2.4e^{-6}$
$\kappa_{0,2}$	$2.5e^{-5}$			$2.47e^{-5}$	$6.4e^{-7}$
$\kappa_{0,3}$	$4e^{-5}$			$3.96e^{-5}$	$1.4e^{-6}$
$\kappa_{0,4}$	$6e^{-5}$			$5.86e^{-5}$	$2.8e^{-6}$
$\kappa_{0,5}$	$1.7e^{-4}$			$1.74e^{-4}$	$6.8e^{-6}$
$\sigma_{1,1}^{10}$	5.5415	N/A		5.30	
$\sigma_{1,2}^{10}$	2.0388	N/A		1.90	
$\sigma_{1,3}^{10}$	4.8541	N/A		5.08	
$\sigma_{1,4}^{10}$	2.8422	N/A		3.34	
$\sigma_{1,5}^{10}$	5.4788	N/A		5.62	

Table 5: Estimates of $\psi = (\{\kappa_{0,i}\}_{i=1}^{5}, \{\delta_i\}_{i=4}^{5}, \{\sigma_{1,i}\}_{i=1}^{5}, \{\sigma_{2,i}\}_{i=4}^{5})$ from 2 years of annual case notification data ($Y^{(1)}$) and vaccinated case notification data ($Y^{(2)}$) with 10% noise. The true values, which were used to generate the data, are denoted by ψ_0.

	ψ_0	$\hat{\psi}$	$SE(\hat{\theta})$
$\kappa_{0,1}$	$3e^{-4}$	$3.00e^{-4}$	$2.6e^{-6}$
$\kappa_{0,1}$	$2.5e^{-5}$	$2.47e^{-5}$	$6.8e^{-7}$
$\kappa_{0,1}$	$4e^{-5}$	$3.96e^{-5}$	$1.5e^{-6}$
$\kappa_{0,1}$	$6e^{-5}$	$5.86e^{-5}$	$3.0e^{-6}$
$\kappa_{0,1}$	$1.7e^{-4}$	$1.76e^{-4}$	$1.4e^{-5}$
δ_4	0.8	0.828	0.054
δ_5	0.6	0.580	0.065
$\sigma_{1,1}$	5.5415	5.60	
$\sigma_{1,2}$	2.0388	2.01	
$\sigma_{1,3}$	4.8541	5.37	
$\sigma_{1,4}$	2.8422	3.53	
$\sigma_{1,5}$	5.4788	5.94	
$\sigma_{2,4}$	0.1007	0.107	
$\sigma_{2,5}$	2.4717	2.80	

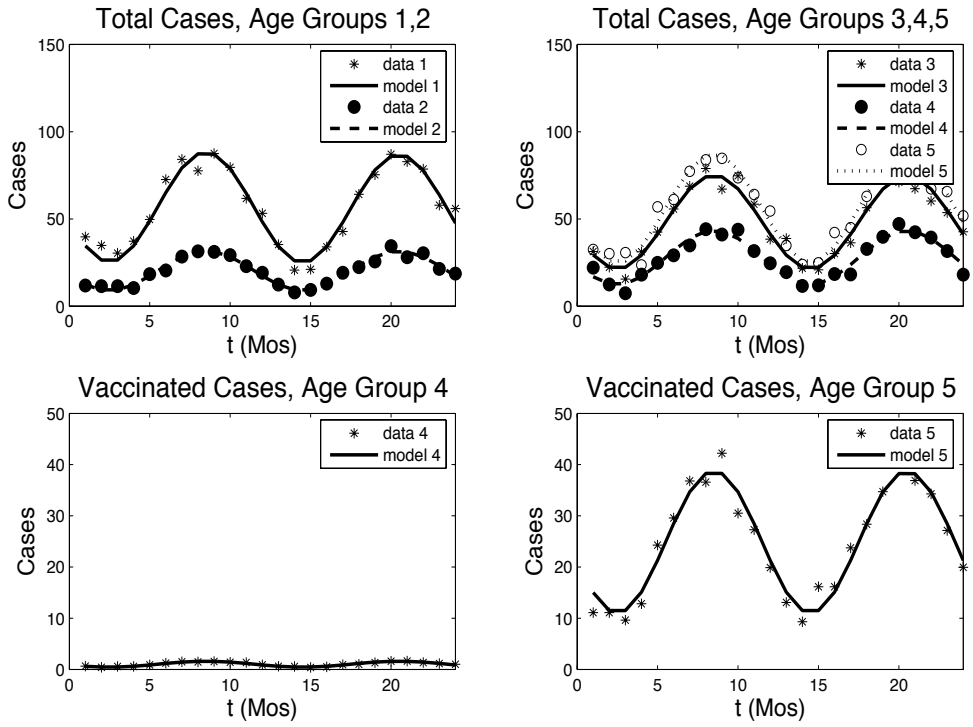

Figure 4: Model fit to 2 years of monthly case notifications ($Y^{(1)}$) and vaccinated cases ($Y^{(2)}$) with 10% noise.

the time-dependent force of infection by a constant $\lambda_i(t) \approx \lambda_i$ since the colonization prevalence is relatively constant in time (Figure 5). Taking the 'true' values here as the mean of $\lambda_i(t)$ for each $i = 1, ..., 5$ over the 2 years of $Y^{(3)}$ data without noise, we can compare these values to estimates of $\{\hat{\lambda}_i\}_{i=1}^5$ from OLS estimation with 2 years of monthly $Y^{(3)}$ with 10% noise (Table 6 and Figure 6). The parameter estimates shown in Table 6 are not as close to their true values as when the mean infection rates and vaccine protection from infection were estimated. However, standard errors are around an order of magnitude less than the parameter values with 10% noise present in the data, indicating that the estimates obtained are reasonably reliable. Given the scarcity of information on both the force of infection and effective contact rates, it is likely advantageous to record colonization data and be able to obtain some quantification of these rates. The routine collection of colonization data, has been considered recently by many public health

departments in light of the development of vaccines which may impact this stage of infection.

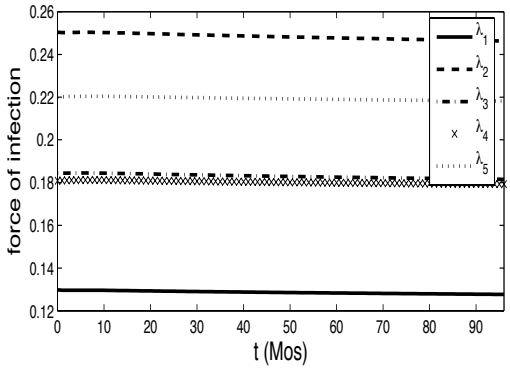

Figure 5: Age-specific forces of infection $\{\lambda_i(t)\}_{i=1}^5$ as a function of time.

We have explicitly included results in which the estimation procedure was successful, and reliable parameter estimates were obtained. However, in numerical experiments not shown here we have tested for the ability to estimate parameters using all forms of data $Y^{(1)}, .., Y^{(4)}$ considered here and were not able to reliably estimate the recovery rate γ. Instead, we explore how other estimated parameters might change in response to the recovery rates changing over the range $\gamma \in [\frac{1}{2}, 2]$. The results in Table 7 suggest that the parameter values and their standard errors are virtually unaffected by changes in the recovery rate in this range, which covers biologically reasonable values. This suggests that although the observations do not contain information on this parameter, it is not likely that fixing it at a slightly erroneous value would deter the calibration of the model.

Table 6: Estimates of $\psi = (\{\lambda_i\}_{i=1}^5, \{\sigma_{3,i}\}_{i=4}^5)$ from 2 years of colonization prevalence data ($Y^{(3)}$) with 0 and 10% noise, where the level of noise is denoted by a superscript.

	ψ_0	$\hat{\psi}^0$	$SE(\hat{\theta}^0)$	$\hat{\psi}^{10}$	$SE(\hat{\theta}^{10})$
Λ_1	0.1286	0.1504	$1.1e^{-4}$	0.1450	0.012
Λ_2	0.2483	0.2905	$2.8e^{-4}$	0.2923	0.035
Λ_3	0.1830	0.2140	$1.5e^{-4}$	0.2252	0.015
Λ_4	0.1802	0.2103	$1.5e^{-4}$	0.2162	0.021
Λ_5	0.2193	0.2558	$1.3e^{-4}$	0.2617	0.018
$\sigma_{3,1}^0$	0	$6.230e^{-6}$			
$\sigma_{3,2}^0$	0	$3.501e^{-5}$			
$\sigma_{3,3}^0$	0	$6.365e^{-5}$			
$\sigma_{3,4}^0$	0	$1.421e^{-5}$			
$\sigma_{3,5}^0$	0	$1.202e^{-5}$			
$\sigma_{3,1}^{10}$	$8.8788e^{-4}$			0.0010	
$\sigma_{3,2}^{10}$	0.0039			0.0049	
$\sigma_{3,3}^{10}$	0.0059			0.0057	
$\sigma_{3,4}^{10}$	0.0023			0.0024	
$\sigma_{3,5}^{10}$	0.0020			0.0024	

Table 7: Estimates of $\psi = (\{\kappa_{0,i}\}_{i=1}^5, \{\delta_i\}_{i=4}^5, \{\sigma_{1,i}\}_{i=1}^5, \{\sigma_{2,i}\}_{i=4}^5)$ from 2 years of monthly total case notification data ($Y^{(1)}$) and vaccinated case data ($Y^{(2)}$) with 10% noise, where the recovery rates have been fixed at $\gamma = \frac{1}{2}, \gamma = 1, \gamma = 2$ for $i = 1, ..., 5$.

		$\gamma = \frac{1}{2}$		$\gamma = 1$		$\gamma = 2$	
	ψ_0	$\hat{\psi}$	$SE(\hat{\theta})$	$\hat{\psi}$	$SE(\hat{\theta})$	$\hat{\psi}$	$SE(\hat{\theta})$
$\kappa_{0,1}$	$3e^{-4}$	$2.99e^{-4}$	$2.5e^{-6}$	$2.97e^{-4}$	$3.0e^{-6}$	$3.00e^{-4}$	$2.5e^{-6}$
$\kappa_{0,2}$	$2.5e^{-5}$	$2.47e^{-5}$	$6.6e^{-7}$	$2.49e^{-5}$	$5.8e^{-7}$	$2.47e^{-5}$	$6.6e^{-7}$
$\kappa_{0,3}$	$4e^{-5}$	$3.95e^{-5}$	$1.4e^{-6}$	$3.93e^{-5}$	$1.4e^{-6}$	$3.96e^{-5}$	$1.5e^{-6}$
$\kappa_{0,4}$	$6e^{-5}$	$5.85e^{-5}$	$2.9e^{-6}$	$6.00e^{-5}$	$2.6e^{-6}$	$5.86e^{-5}$	$2.9e^{-6}$
$\kappa_{0,5}$	$1.7e^{-4}$	$1.76e^{-4}$	$1.4e^{-5}$	$1.71e^{-4}$	$1.2e^{-5}$	$1.76e^{-4}$	$1.4e^{-5}$
δ_4	0.8	0.828	0.053	0.797	0.051	0.828	0.053
δ_5	0.6	0.580	0.063	0.601	0.056	0.580	0.063
$\sigma_{1,1}$	5.5415	5.448		6.43		5.415	
$\sigma_{1,2}$	2.0388	1.9491		1.71		1.950	
$\sigma_{1,3}$	4.8541	5.221		4.88		5.223	
$\sigma_{1,4}$	2.8422	3.426		3.08		3.428	
$\sigma_{1,5}$	5.4788	5.775		5.08		5.773	
$\sigma_{2,4}$	0.1007	0.1019		0.117		0.1019	
$\sigma_{2,5}$	2.4717	2.726		2.36		2.725	

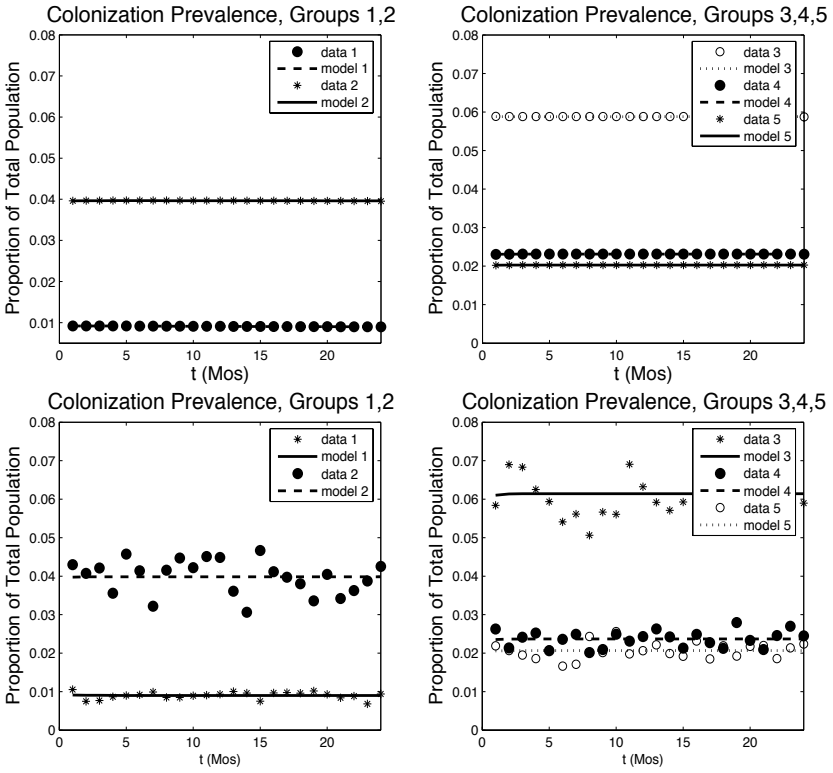

Figure 6: Model fit to 2 years of monthly colonization prevalence data ($Y^{(4)}$) with 0% noise (top panels), and 10% noise (bottom panels).

6 Assessment of vaccine programs

To assess the effectiveness of a given vaccine program once it has been implemented, we would like to be able to measure the protection induced by the vaccine. The protection from infection factor $1 - \delta_i$ or what is typically referred to as vaccine efficacy, can differ in practice from what is reported in vaccine efficacy studies. Therefore, it is important to discuss how to effectively estimate this parameter from surveillance data. In the case of pneumococcal infections, the colonized state is particularly relevant in disease dynamics, and the ability of some novel vaccines to affect this state is both controversial and promising. Experts have not agreed whether the conjugate vaccine has had a significant impact on colonization, nor if the impact would be advantageous or harmful. Thus, measuring the reduction of asymptomatic carriage due to vaccination ν_1, ν_2 in the younger age classes is crucial to study the impact of recently developed vaccines, and those that are currently in development.

As in Section 5, we use reports of vaccinated cases in the OLS procedure to estimate the protection from infection, and we now estimate this parameter for the younger and older age classes $\{\delta_i\}_{i=1,2,4,5}$, shown in the third and fourth columns of Table 8. The seventh and eighth columns contain estimates of ν_1, ν_2 from records of colonization prevalence in vaccinated individuals ($Y^{(4)}$), corresponding to nasal swabs in which vaccination information has been recorded. Since the estimates are relatively close to their true values, and the standard errors are small compared to the parameter values, these results suggest that data of these forms contain enough information on the parameters to reliably estimate them. In the last two columns of Table 8, we see the results of estimating these parameters simultaneously from these data. The parameter estimates are relatively close to their true values and the standard errors for some of the parameters indicate that the estimates are still reasonably reliable if they are estimated simultaneously. Thus, information on one aspect of vaccine protection is not needed to estimate the other, and the unknown parameters may be reliably estimated simutaneously.

Collection of vaccination information involves additional resources and requires greater effort, and therefore may not always be a feasible option for some public health offices. Therefore, we have explored the possibility of estimating these parameters simultaneously from case notification and colonization data when vaccination information was not given for one type of data (not shown). Unfortunately, the parameter estimates do not converge to their true values, even when five years of monthly data is used in the

Table 8: Estimates of $\psi = (\{\delta_i\}_{i=1,2,4,5}, \{\sigma_{2,i}\}_{i=1,2,3,4,5})$ from 2 years of monthly vaccinated case notification data ($Y^{(2)}$) with 10% noise, shown in columns 3 and 4. Columns 5 and 6 contain estimates of $\psi = (\{\nu_i\}_{i=1,2}, \{\sigma_{4,i}\}_{i=1,2,3,4,5})$ from 2 years of monthly vaccinated colonization prevalence data ($Y^{(4)}$) with 10% noise. Estimates of all parameters simultaneously $(\{\delta_i\}_{i=1,2,4,5}, \{\nu_i\}_{i=1,2}, \{\sigma_{2,i}\}_{i=1,2,3,4,5}, \{\sigma_{4,i}\}_{i=1,2,3,4,5})$, from $Y^{(2)}$ and $Y^{(4)}$ are shown in columns 7 and 8.

		$Y^{(2)}$		$Y^{(4)}$		$Y^{(2)}, Y^{(4)}$	
	ψ_0	$\hat{\psi}$	$SE(\hat{\theta})$	$\hat{\psi}$	$SE(\hat{\theta})$	$\hat{\psi}$	$SE(\hat{\theta})$
δ_1	0.2	0.2004	0.00033			0.2039	0.00030
δ_2	0.4	0.4022	0.015			0.4024	0.017
δ_4	0.8	0.7937	0.023			0.7941	0.024
δ_5	0.6	0.6093	0.022			0.6096	0.023
ν_1	0.3			0.2952	0.016	0.2938	0.0011
ν_2	0.5			0.4981	0.00055	0.5001	0.0049
$\sigma_{2,1}$	0.2224	0.0429				0.04412	
$\sigma_{2,2}$	0.1062	0.1029				0.1061	
$\sigma_{2,3}$	0.09391	0.0918				0.09432	
$\sigma_{2,4}$	0.05609	0.0430				0.04458	
$\sigma_{2,5}$	1.2660	1.4061				1.4776	
$\sigma_{4,1}$	0.006761			0.001410		0.001558	
$\sigma_{4,2}$	0.00141			0.001836		0.002030	
$\sigma_{4,3}$	0.004100			0.0001781		0.001969	
$\sigma_{4,4}$	0.002051			0.0002552		0.002821	
$\sigma_{4,5}$	0.02165			0.003595		0.003975	

estimation process. This suggests that an emphasis on the collection of vaccination information would be necessary should a vaccination program need to be assessed.

Table 9: Mean infection rates used to generate data with the 14-age-class model for use in the model comparison demonstration.

age group i	mean infection rate $\kappa_{0,i}$
(0,2 mos]	$2e^{-4}$
(2,4 mos]	$3.2e^{-4}$
(4,6 mos]	$2e^{-3}$
(6,24 mos]	$3.5e^{-4}$
(2,5 yrs]	$6e^{-5}$
(5,10 yrs]	$3e^{-5}$
(10,15 yrs]	$1.5e^{-5}$
(15,50 yrs]	$4.8e^{-5}$
(50,65 yrs]	$6.3e^{-5}$
(65,70 yrs]	$1.8e^{-4}$
(70,75 yrs]	$1.9e^{-4}$
(75,80 yrs]	$2e^{-4}$
(80,85 yrs]	$2.1e^{-4}$
(85,∞)	$1.9e^{-4}$

7 Age class refinement

Additional information accompanying a data set, such as age, is usually beneficial, although the level of detail which should be incorporated to the model is not always clear. The incorporation of too much information in a model can hinder computations and unnecessarily increase complexity. Here we present a discussion of the level of refinement that significantly improves agreement between the mathematical model and a set of observations. For this illustration, data has been generated by a model made up of 14 distinct age classes. The age classes are: (0,2 mos], (2,4 mos] (4,6 mos], (6,24 mos], (2,5 yrs], (5,10 yrs], (10,15], (15,50], (50,65], (65,70], (70,75], (75,80], (80,85], (85,∞). The methods used here apply to 'nested' models so this is the most discretized version of age groups, and all other models will consider only aggregations of these age groups.

Contained in Table 9 are the 'true' values for the mean infection rates which were used to generate the 'data'. The infection rates among the first 4 age classes differ the most although the intervals are the of the shortest length; thus it is expected that the age-dependence in this range is important. In contrast, the infection rates among the last five age classes are

relatively similar. If we consider these age classes as distinct, we will obtain an 'improved fit' indicated by a smaller residual. However, whether this reduction indicates a significant improvement requires the use of the statistical test described in Section 4.3.

Consider the model which aggregates the ages into the following age classes: (0,2 yrs], (2,15 yrs], (15,50], (50,65], (65, ∞). Fitting this model to data which has been aggregated according to the constraint $H\theta = 0$ where H is an 9 x 14 matrix, of which the first 5 rows are given by

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & -1 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & \cdots & 0 \end{pmatrix}.$$

Thus, the constraints are $\kappa_{0,1} = \kappa_{0,2} = \kappa_{0,3} = \kappa_{0,4}$, $\kappa_{0,5} = \kappa_{0,6} = \kappa_{0,7}$, and $\kappa_{0,10} = ...\kappa_{0,14}$. Imposing this constraint we obtain the results in the fit shown in Figure 7. The model does not appear to agree well with the data, and it is clear that some information has been lost. However, this would not be obvious if the 5 age group model fit is plotted with the data that has also been grouped into the same age classes, as seen in Figure 8. The residual calculated from the fit seen in Figure 7 is used in the model comparison statistic. We let $J^{14}_{60}(y, \hat{\theta}^{14})$ denote the residual from the 14 age-class model fit to infection data and $J^{5}_{60}(y, \hat{\theta}^{5})$ denote that from the model with 5 age classes fit to data. The test statistic is then

$$\hat{U}^{9}_{60}(y) = \frac{60 * (J^{5}_{60}(y, \hat{\theta}^{5}) - J^{14}_{60}(y, \hat{\theta}^{14})}{J^{14}_{60}(y, \hat{\theta}^{14})} \tag{47}$$

which is compared to a χ^2 distribution with 9 degrees of freedom. As reported in Table 10, even with a significance level of 99.5%, we would reject the null hypothesis in this case. The null hypothesis is that the true parameters θ_0 lie within the constrained parameter space $\Theta_H = \{\theta \in \Theta | H\theta = 0\}$. Thus, this suggests that the additional information provided by considering a finer age discretization does result in a significant improvement in the model agreement to the 'observations'. That is, a researcher would be reasonably motivated to include the additional detail in the model when studying these data.

Table 10 contains the relevant quantities for determining an appropriate level of age structure refinement. The age classes have been grouped together for the intervals shown in the top row. Comparing the first with

Table 10: Results of models with varying age structure fitted to 'data' generated from 14 age classes. $\alpha = 0.0005$ for all values of τ reported below, and U_{60} was calculated with $J_{60}^{14}(y, \theta^{14}) = 2,083.7$

Aggregated age groups	(0,2 y] (2,15] (65,∞)	(2,15] (65,∞)	(65, ∞)
r: $\chi^2(r)$	9	6	4
τ	29.67	24.10	20.00
J_{60}	9,475.1	3,093.7	2,244.1
U_{60}^r	212.8	29.08	4.62

the second column of numerical values we find there is a marked decrease in residual as the age structure is included in the younger ages. It is unclear from inspection whether the residual in the case of the second column is sufficiently close to that of the fully discretized (14 age-class) model to justify the grouping in ages (2,15] and over 65. However, the test statistic is greater than the threshold value for the very high significance level used here, indicating that even in this case, we should reject the null hypothesis. That is, the improvement in the data fit here is significant and the additional detail in these age classes should be incorporated. When only the age groups over 65 have been aggregated, the residual appears close to the fully discretized model and the statistic supports the null hypothesis. In fact, even for a low significance level of $\alpha = 0.25$, $\hat{U} > \tau = 5.39$. In this case, one would be justified in ignoring the age intervals over 65.

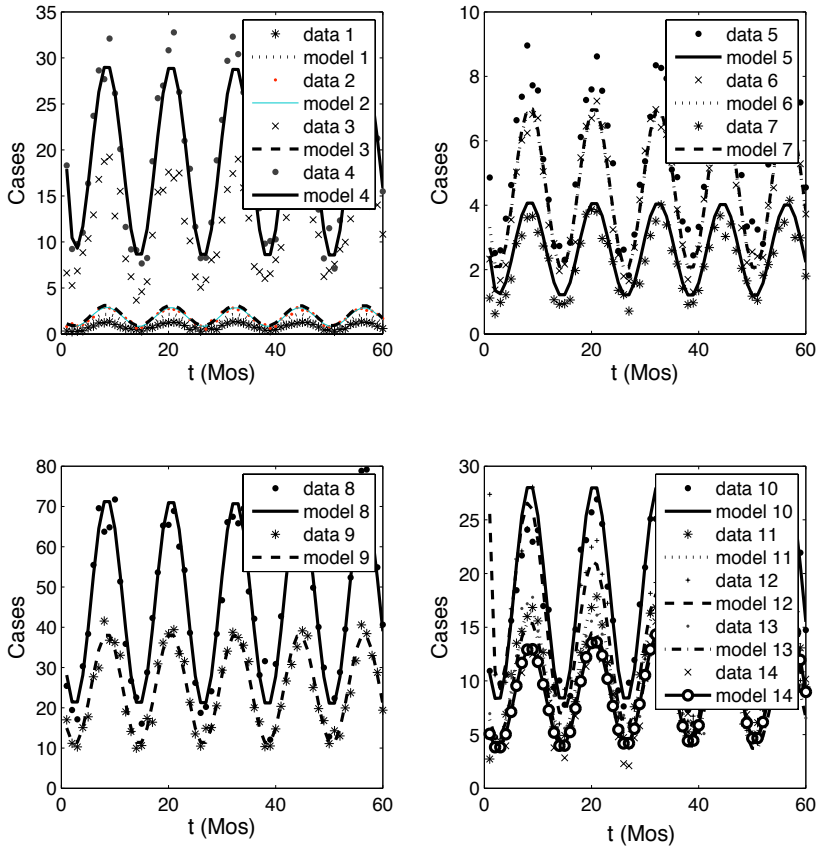

Figure 7: Model with 5 age classes plotted with infection data grouped into 14 age classes.

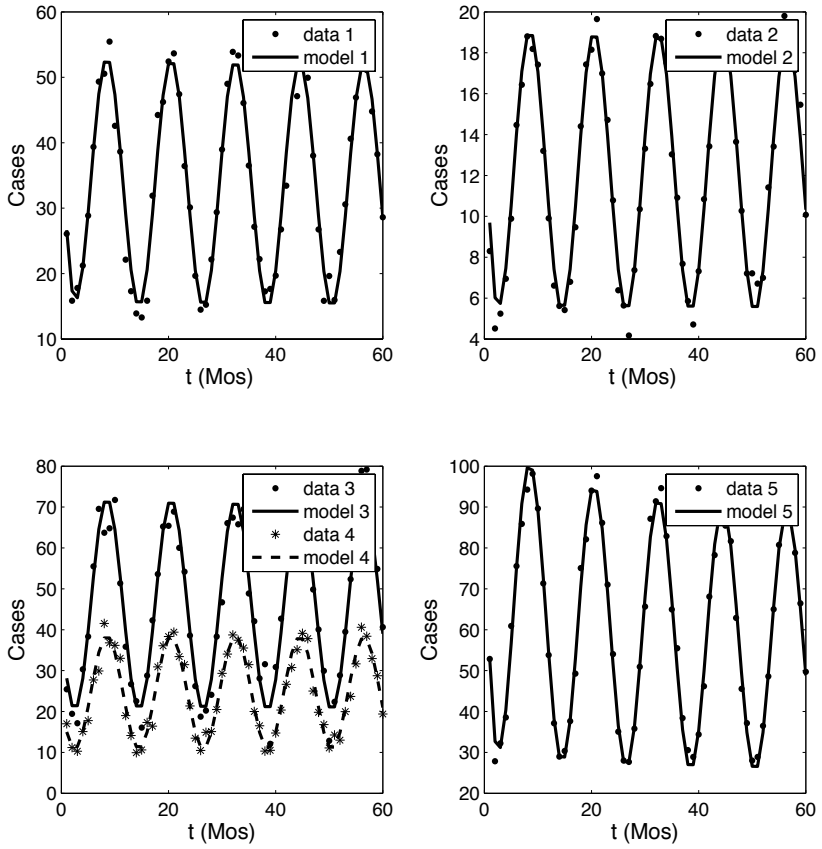

Figure 8: Model with 5 age classes plotted with infection data grouped into 5 age classes.

8 Conclusions

In pneumococcal disease dynamics and many other infectious diseases, age structure plays a crucial role. When prescribing a vaccination strategy for a particular population, these effects should be explicitly incorporated. How to effectively immunize a population against most infections remains a challenge. Public health departments need to be able to use the information from surveillance data to describe as completely as possible the landscape of disease dynamics, including any age-dependent behavior, in a population. Further, once a strategy has been implemented, its impact is reassessed via surveillance data. Here, we have illustrated the use of inverse problem methods as a means to effectively use surveillance data with mathematical modeling. Specifically, we have illustrated the calibration of an age-structured model of pneumococcal diseases including vaccination to a population. We have also described the use of this calibrated model to assess the impact of a given vaccination policy. Further, we discussed how an appropriate level of aggregation in the mathematical model may be determined.

Before a novel vaccine policy has been implemented, the calibration of the mathematical model can be used to understand the dynamics of the infections, the potential impact of vaccination, and to make predictions under certain control scenarios. We have illustrated that case notification data, together with vaccination information, allows for the estimation of infection rates and the efficacy of any vaccine that is currently implemented in the population. The polysaccharide vaccine (PPV23) has been licensed in most developed countries since 1983, and it is necessary to account for this in the model calibration. We have shown that if annual data are not available for sufficiently many years to estimate these parameters, it is comparable to use more frequent data, say monthly recordings. While we were not able to estimate the effective contact rates, if the colonization prevalence appears to be relatively constant over a period of time, it is reasonable to approximate the force of infection by a constant, and to estimate it from these data. This is particularly advantageous, in light of the historic lack of information on this parameter, which has such an important role in governing the overall epidemiological dynamics.

In the assessment of a vaccination strategy, a calibrated model can be used to estimate only parameters that would be changed. Provided there is information available on vaccination rates and the duration of protection of the vaccines implemented, the only remaining, however critical, parameters quantify the protection from infection and colonization. As was illustrated in the model calibration studies, the total cases occurring in vaccinated indi-

viduals allows for the quantification of the protection from infection induced by the vaccine. If vaccination history is recorded for individuals from which nasal swabs have been taken we can also estimate the protection from colonization. Although these additional records require more resources and efforts, our studies indicate that the information is necessary to effectively evaluate a vaccine policy.

The asymptomatic colonized state, which precedes invasive infection, is important in pneumococcal disease dynamics. The controversial impact of the conjugate vaccine and the potential impact of other developmental vaccines on this state have prompted discussion of proposed uses for regularly collected nasal swab data. This would give an indication of the prevalence of colonization. Here we have demonstrated that such information could be quite useful, as it allows for the estimation of the force of infection, a parameter that is otherwise unavailable. In addition, with the inclusion of vaccination history of swabbed individuals, the effect of immunization on this state would be quantifiable. To monitor whether vaccine policies have any impact on colonization - a controversial and pertinent issue - the collection of nasal swabs along with immunization status of individuals is necessary, and should be strongly emphasized.

With many observations, there typically comes additional information, such as age, geographical location, and strain or serotype information. Whether to include this information in theoretical studies and at what level of detail is not always clear by inspection of the data. We have described via a model comparison statistic how one might determine a suitable refinement of age groups given a set of age-stratified data. Such a tool is useful to elucidate important aspects of disease dynamics, and perhaps underlying mechanisms driving disease persistence that should be targeted by control strategies. Additionally, this test provides a safeguard against unnecessary computational challenges.

Acknowledgements

This research was supported in part by the National Institute of Allergy and Infectious Diseases under grant 9R01AI071915-05, the National Science Foundation under grant DMS-0502349, an Achievement Rewards for College Scientists Foundation Scholarship, generously donated by Ralph and Sandra Matteucci, and a competitive Research Grant awarded by the Graduate and Professional Students Association at Arizona State University. Finally, the first and second authors are thankful for the hospitality of the Radon

Institute for Computational and Applied Mathematics (RICAM) in Linz, Austria, during a visit in which the writing of this manuscript was completed.

References

[1] H. R. Altuzarra, B. M. T. Valenzuela, A. O. Trucco, S. J. Inostroza, S. P. Granata, and V. J. Fleiderman. *Nasal carriage of Streptococcus pneumoniae in elderly subjects according to vaccination status*, Revista medica de chile, **135** (2007), 160–166.

[2] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, London, 1989.

[3] R. Austrian. *Pneumococcus: the first one hundred years*, Reviews in Infectious Diseases, **3** (1981), 183–189.

[4] H. T. Banks, M. Davidian, J. R. Samuels, Jr. and K. L. Sutton. *An inverse problem statistical methodology summary*, Center for Research in Scientific Computation Technical Report, CRSC-TR08-01, January, 2008, North Carolina State University.

[5] H.T. Banks, J. E. Banks, L. K. Dick, and J. D. Stark. *Estimation of dynamic rate parameters in insect populations undergoing sublethal expoure to pesticides* CRSC-TR05-22, NCSU, May, 2005; Bulletin of Mathematical Biology, **69** (2007), 2139–2180.

[6] H. T. Banks and B. G. Fitzpatrick. *Estimation of growth rate distributions in size-structured population models*, CAMS Tech. Rep. 90-2, University of Southern California, January, 1990; Quarterly of Applied Mathematics, **49** (1991), 215–235.

[7] H. T. Banks and B. G. Fitzpatrick. *Statistical methods for model comparison in parameter estimation problems for distributed systems* , Journal of Mathematical Biology, **28** (1990), 501–527.

[8] H. T. Banks and K. Kunisch. *Estimation tecniques for distributed parameter systems*, Birkhauser, Boston, 1989.

[9] H. T. Banks, L. W. Botsford, F. Kappel, and C. Wang. *Modeling and estimation in size structured population models*, LCDS-CCS Report 87-13, Brown University; *Proceedings 2nd Course on Mathematical*

Ecology, (Trieste, December 8-12, 1986) World Press (1988), Singapore, 521–541.

[10] F. Brauer and C. Castillo-Chavez *Mathematical Models in Population Biology and Epidemiology*, Springer-Verlag, New York, 2001.

[11] A. E. Bridy-Pappas, M. B. Margolis, K. J. Center and Isaacman. *Streptococcus pneumoniae: description of the pathogen, disease epidemiology, treatment, and prevention*, Pharmacotherapy, **25** (2005), 1193–1212.

[12] Communicable Diseases Australia, National Notifiable Diseases Surveillance System. http://www9.health.gov.au/cda/Source/CDA-index.cfm

[13] M. Davidian and D. Giltinan. *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London, 1998.

[14] A. R. Gallant. *Nonlinear Statistical Models*, Wiley, New York, 1987.

[15] H. W. Hethcote. *An age-structured model for pertussis transmission*, Mathematical Biosciences, **145** (1997), 89–136.

[16] H. W. Hethcote. *Simulations of pertussis epidemiology in the United States: effects of adult booster vaccinations*, Mathematical Biosciences, **158** (1999), 47–73.

[17] H. W. Hethcote, P. Horby and P. McIntyre. *Using computer simulations to compare pertussis vaccination strategies in Australia*, Vaccine, **22** (2004), 2181–2191.

[18] R. I. Jennrich. *Asymptotic properties of non-linear least squares estimators.*, Annals of Mathematical Statistics, **40** (1969), 633–643.

[19] E. N. Janoff and J. B. Rubins. *Invasive pneumococcal disease in the immunocompromised host.*, Microbial Drug Resistance, **3** (1997), 215–232.

[20] D. M. Musher, *Streptococcus pneumoniae*, in "Mandell, Douglas, and Bennett's principles and practice of infectious diseases" (eds. G. E. Mandell, J. E. Bennett and R. Dolin), Churchill Livingstone, (2000), 2128–2144.

[21] P. Roche and V. Krause. *Invasive pneumococcal disease in Australia, 2001*, Communicable Diseases Intelligence, **26** (2002), 505–519.

[22] P. Roche, V. Krause, R. Andrews, L. Carter, D. Coleman, H. Cook, M. Counahan, C. Giele, R. Gilmore, S. Hart and R. Pugh. *Invasive pneumococcal disease in Australia, 2002*, Communicable Diseases Intelligence, **27** (2003), 466–477.

[23] P. Roche, V. Krause, M. Bartlett, D. Coleman, H. Cook, M. Counahan, C. Davis, L. Del Fabbro, C. Geile, R. Gilmore, R. Kampen and M. Young. *Invasive pneumococcal disease in Australia, 2003*, Communicable Diseases Intelligence, **28** (2004), 441–454.

[24] G. A. F. Seber and J. Wild. *Nonlinear regression*, Wiley, New York, 1989.

[25] K. L. Sutton, H. T. Banks and C. Castillo-Chávez. *Estimation of invasive pneumococcal disease dynamic parameters and the impact of conjugate vaccination in Australia*, Mathematical Biosciences and Engineering, **5** (2008), 176–2004.

[26] K. L. Sutton, H. T. Banks and C. Castillo-Chávez. *An age-structured model for pneumococcal infection with vaccination*, Center for Research in Scientific Computation Technical Report, CRSC-TR08-13, September, 2008, North Carolina State University.

[27] R. K. Syrjanen, T. M. Kilpi, T. H. Kaijalainen, E. E. Herva and A. K. Takala. *Nasopharyngeal carriage of Streptococcus pneumoniae in Finnish children younger than 2 years old* Journal of Infectious Diseases, **184** (2001), 145–149.

[28] World Health Organization, *Pneumococcal vaccines*, WHO Weekly Epidemiological Record, **79** (1999), 177–183.