MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **6**, Number **2**, April **2009**

doi:10.3934/mbe.2009.6.261

pp. **261–282**

THE ESTIMATION OF THE EFFECTIVE REPRODUCTIVE NUMBER FROM DISEASE OUTBREAK DATA

Ariel Cintrón-Arias

Center for Research in Scientific Computation
Center for Quantitative Sciences in Biomedicine
North Carolina State University, Raleigh, NC 27695, USA

Carlos Castillo-Chávez

Department of Mathematics and Statistics
Arizona State University, P.O. Box 871804, Tempe, AZ 85287-1804, USA

Luís M. A. Bettencourt

Theoretical Division, Mathematical Modeling and Analysis (T-7)
Los Alamos National Laboratory, Mail Stop B284, Los Alamos, NM 87545, USA

Alun L. Lloyd and H. T. Banks

Center for Research in Scientific Computation
Biomathematics Graduate Program
Department of Mathematics
North Carolina State University, Raleigh, NC 27695, USA

Abstract. We consider a single outbreak susceptible-infected-recovered (SIR) model and corresponding estimation procedures for the effective reproductive number $\mathcal{R}(t)$. We discuss the estimation of the underlying SIR parameters with a generalized least squares (GLS) estimation technique. We do this in the context of appropriate statistical models for the measurement process. We use asymptotic statistical theories to derive the mean and variance of the limiting (Gaussian) sampling distribution and to perform post statistical analysis of the inverse problems. We illustrate the ideas and pitfalls (e.g., large condition numbers on the corresponding Fisher information matrix) with both synthetic and influenza incidence data sets.

1. Introduction.

The transmissibility of an infection can be quantified by its basic reproductive number $\mathcal{R}_0$, defined as the mean number of secondary infections seeded by a typical infective into a completely susceptible (naïve) host population [1, 19, 26]. For many simple epidemic processes, this parameter determines a threshold: whenever $\mathcal{R}_0 > 1$, a typical infective gives rise, on average, to more than one secondary infection, leading to an epidemic. In contrast, when $\mathcal{R}_0 < 1$, infectives typically give rise, on average, to less than one secondary infection, and the prevalence of infection cannot increase.

2000 *Mathematics Subject Classification.* Primary: 62G05, 93E24, 49Q12, 37N25; Secondary: 62H12, 62N02.

Key words and phrases. effective reproductive number, basic reproduction ratio, reproduction number, $\mathcal{R}$, $\mathcal{R}(t)$, $\mathcal{R}_0$, parameter estimation, generalized least squares, residual plots.

The first author was in part supported by NSF under Agreement No. DMS-0112069, and by NIH Grant Number R01AI071915-07.

Owing the natural history of some infections, transmissibility is better quantified by the *effective*, rather than the basic, reproductive number. For instance, exposure to influenza in previous years confers some cross-immunity [16, 22, 32]; the strength of this protection depends on the antigenic similarity between the current year's strain of influenza and earlier ones. Consequently, the population is non-naïve, and so it is more appropriate to consider the effective reproductive number $\mathcal{R}(t)$, a time-dependent quantity that accounts for the population's reduced susceptibility.

Our goal is to develop a methodology for the estimation of $\mathcal{R}(t)$ that also provides a measure of the uncertainty in the estimates. We apply the proposed methodology in the context of annual influenza outbreaks, focusing on data for influenza A (H3N2) viruses, which were, with the exception of the influenza seasons 2000–01 and 2002–03, the dominant flu subtype in the United States (US) over the period from 1997 to 2005 [12, 36].

The estimation of reproductive numbers is typically an indirect process because some of the parameters on which these numbers depend are difficult, if not impossible, to quantify directly. A commonly used indirect approach involves fitting a model to some epidemiological data, providing estimates of the required parameters.

In this study we estimate the effective reproductive number by fitting a deterministic epidemiological model employing a generalized least squares (GLS) estimation scheme to obtain estimates of model parameters. Statistical asymptotic theory [18, 34] and sensitivity analysis [17, 33] are then applied to give approximate sampling distributions for the estimated parameters. Uncertainty in the estimates of $\mathcal{R}(t)$ is then quantified by drawing parameters from these sampling distributions, simulating the corresponding deterministic model and then calculating effective reproductive numbers. In this way, the sampling distribution of the effective reproductive number is constructed at any desired time point.

The statistical methodology provides a framework within which the adequacy of the parameter estimates can be formally assessed for a given data set. We discuss the use of residual plots as a diagnostic for the estimation, highlighting the problems that arise when the assumptions of the statistical model underlying the estimation framework are violated.

This manuscript is organized as follows: In Section 2 the data sets are introduced. A single-outbreak deterministic model is introduced in Section 3. Section 4 introduces the least squares estimation methodology used to estimate values for the parameters and quantify the uncertainty in these estimates. Our methodology for obtaining estimates of $\mathcal{R}(t)$ and its uncertainty is also described. Use of these schemes is illustrated in Section 5, in which they are applied to synthetic data sets. Section 6 applies the estimation machinery to the influenza incidence data sets. We conclude with a discussion of the methodologies and their application to the data sets.

2. Longitudinal incidence data.

Influenza is one of the most significant infectious diseases of humans, as witnessed by the 1918 "Spanish flu" pandemic, during which 20% to 40% of the worldwide population became infected. At least 50 million deaths resulted, with 675,000 of these occurring in the US [37]. The impact of flu is still significant during inter-pandemic periods: the Centers for Disease Control and Prevention (CDC) estimate that between 5% and 20% of the US population becomes infected annually [12]. These annual flu outbreaks lead to an average

Table 1. Number of tested specimens and influenza isolates during several annual outbreaks in the US [12].

Season	Total number of tested specimens	Number of A(H1N1) & A(H1N2) isolates	Number of A(H3N2) isolates	Number of B isolates
1997–98	99,072	6	3,241	102
1998–99	102,105	30	2,607	3,370
1999–00	92,403	132	3,640	77
2000–01	88,598	2,061	66	4,625
2001–02	100,815	87	4,420	1,965
2002–03	97,649	2,228	942	4,768
2003–04	130,577	2	7,189	249
2004–05	157,759	18	5,801	5,799
Mean	108,622	571	3,488	2,619

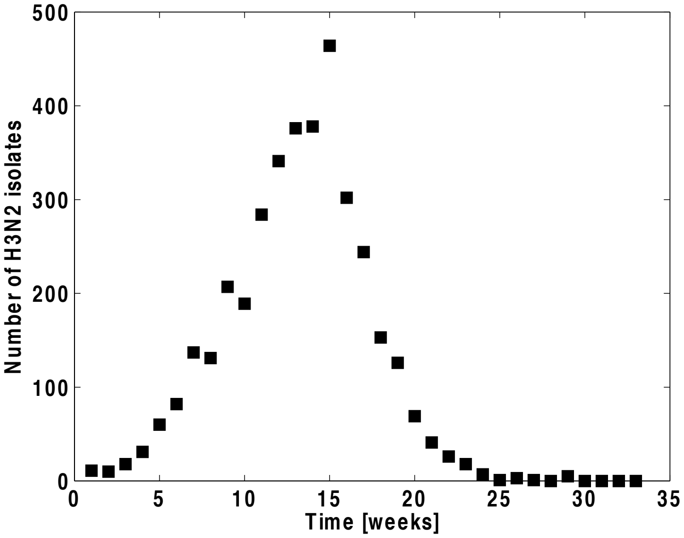

Figure 1. Influenza isolates reported by the CDC in the US during the 1999–00 season [12]. The number of H3N2 cases (isolates) is displayed as a function of time. Time is measured as the number of weeks since the start of the year's flu season. For the 1999–00 flu season, week number one corresponds to the fortieth week of the year, falling in October.

of 200,000 hospitalizations (mostly involving young children and the elderly) and mortality that ranges between about 900 and 13,000 deaths per year [36].

The Influenza Division of the CDC reports weekly information on influenza activity in the US from calendar week 40 in October through week 20 in May [12], the period referred to as the influenza season. Because the influenza virus exhibits a high degree of genetic variability, data is not only collected on the number of cases but also on the types of influenza viruses that are circulating. A sample of viruses isolated from patients undergoes antigenic characterization, with the type, subtype and, in some instances, the strain of the virus being reported [12].

The CDC acknowledges that, while these reports may help in mapping influenza activity (whether or not it is increasing or decreasing) throughout the US, they often do not provide sufficient information to calculate how many people became ill with influenza during a given season. This is true especially in light of measurement uncertainty, e.g., underreporting, longitudinal variability in reporting procedures, etc. Indeed, the sampling process that gives rise to the tested isolates is not sufficiently standardized across space and time, and results in variabilities in measurements that are difficult to quantify. We return to discuss this point later in this paper.

Despite the cautionary remarks by the CDC, we use such isolate reports as illustrative data sets to which one can apply proposed estimation methodologies. The data sets do, in fact, represent typical data sets available to modelers for many disease progression scenarios. Interpretation of the results, however, should be mindful of the issues associated with the data. For the influenza data we have chosen, the total number of tested specimens and isolates through various seasons are summarized in Table 1. It is observed that H3N2 viruses predominated in most seasons with the exception of 2000–01 and 2002–03. Consequently, we focus our attention on the H3N2 subtype. Fig. 1 depicts the number of H3N2 isolates reported over the 1999–00 influenza season.

3. **Deterministic single-outbreak SIR model.** The model that we use is the standard susceptible-infected-recovered (SIR) model (see, for example, [1, 8]). The state variables $S(t)$, $I(t)$, and $X(t)$ denote the number of people who are susceptible, infected, and recovered, respectively, at time t. It is assumed that newly infected individuals immediately become infectious and that recovered individuals acquire permanent immunity. The influenza season, lasting nearly thirty-two weeks [12], is short compared to the average lifespan, so we ignore demographic processes (births and deaths) as well as disease-induced fatalities and assume that the total population size remains constant. The model is given by the set of nonlinear differential equations

$$\frac{dS}{dt} = -\beta S\frac{I}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta S\frac{I}{N} - \gamma I \tag{2}$$

$$\frac{dX}{dt} = \gamma I. \tag{3}$$

Here β is the transmission parameter and γ is the (per-capita) rate of recovery, the reciprocal of which gives the average duration of infection. Observe that one of the differential equations is redundant because the three compartments sum to the constant population size: $S(t) + I(t) + X(t) = N$. We choose to track $S(t)$ and $I(t)$. The initial conditions of these state variables are denoted by $S(t_0) = S_0$ and $I(t_0) = I_0$.

Equation (2) for the infective population can be rewritten as

$$\frac{dI}{dt} = \gamma(\mathcal{R}(t) - 1)I, \tag{4}$$

where $\mathcal{R}(t) = \frac{S(t)}{N}\mathcal{R}_0$ and $\mathcal{R}_0 = \beta/\gamma$. $\mathcal{R}(t)$ is known as the effective reproductive number, while $\mathcal{R}_0$ is known as the basic reproductive number. We have that $\mathcal{R}(t) \leq \mathcal{R}_0$, with the upper bound—the basic reproductive number—only being achieved when the entire population is susceptible.

We note that $\mathcal{R}(t)$ is the product of the per-infective rate at which new infections arise and the average duration of infection, and so the effective reproductive number gives the average number of secondary infections caused by a single infective, at a given susceptible fraction. The prevalence of infection increases or decreases according to whether $\mathcal{R}(t)$ is greater than or less than one, respectively. Because there is no replenishment of the susceptible pool in this SIR model, $\mathcal{R}(t)$ decreases over the course of an outbreak as susceptible individuals become infected.

4. **Estimation scheme.**

To calculate $\mathcal{R}(t)$, one needs to know the two epidemiological parameters β and γ, as well as the number of susceptibles $S(t)$ and the population size N. As mentioned before, difficulties in the direct estimation of β, whose value reflects the rate at which contacts occur in the population and the probability of transmission occurring when a susceptible and an infective meet, and direct estimation of $S(t)$ preclude direct estimation of $\mathcal{R}(t)$. As a result, we adopt an indirect approach, which proceeds by first finding the parameter set for which the model has the best agreement with the data and then calculating $\mathcal{R}(t)$ by using these parameters and the model-predicted time course of $S(t)$. Simulation of the model also requires knowledge of the initial values, S_0 and I_0, which must also be estimated.

Although the model is framed in terms of the prevalence of infection $I(t)$, the time-series data provides information on the weekly incidence of infection, which, in terms of the model, is given by the integral of the rate at which new infections arise over the week: $\int \beta S(t)I(t)/N\,dt$. We observe that the parameters β and N only appear (both in the model and in the expression for incidence) as the ratio β/N, precluding their separate estimation. Consequently we need only estimate the value of this ratio, which we denote by $\tilde{\beta} = \beta/N$.

We employ inverse problem methodology to obtain estimates of the vector $\theta = (S_0, I_0, \tilde{\beta}, \gamma) \in \mathbb{R}^p = \mathbb{R}^4$ by minimizing the difference between the model predictions and the observed data, according to a generalized least squares (GLS) criterion. In what follows, we refer to θ as the parameter vector, or simply as the parameter, in the inverse problem, even though some of its components are initial conditions rather than parameters, of the underlying dynamic model.

4.1. **Generalized Least Squares (GLS) estimation.**

The least squares estimation methodology is based on a *statistical model* for the observation process (referred to as the case-counting process) as well as the *mathematical model.* As is standard in many statistical formulations, it is assumed that our known model, together with a particular choice of parameters (the “true” parameter vector, written as θ_0) exactly describes the epidemic process, but that the n observations $\{Y_j\}_{j=1}^n$ are affected by random deviations (e.g., measurement errors) from this underlying process. More precisely, it is assumed that

$$Y_j = z(t_j;\theta_0) + z(t_j;\theta_0)^\rho \epsilon_j \qquad \text{for } j = 1,\ldots,n \tag{5}$$

where $z(t_j;\theta_0)$ denotes the weekly incidence given by the model under the true parameter, θ_0, and is defined by the integral

$$z(t_j;\theta_0) = \int_{t_{j-1}}^{t_j} \tilde{\beta} S(t;\theta_0) I(t;\theta_0)\,dt. \tag{6}$$

Here t_0 denotes the time at which the epidemic observation process started and the weekly observation time points are written as $t_1 < \cdots < t_n$.

We remark that the choice of a particular statistical model (i.e., the error model for the observation process) is often a difficult task. While one can never be certain of the correctness of one's choice, there are post-inverse problem quantitative methods (e.g., involving residual plots) that can be effectively used to investigate this question; see the discussions and examples in [3]. A major goal of this paper is to present and illustrate use of such ideas and techniques in the context of surveillance data modeling.

The "errors" ϵ_j (note that the total measurement errors $\tilde{\epsilon}_j = z(t_j;\theta_0)^\rho \epsilon_j$ are model-dependent) are assumed to be independent and identically distributed (*i.i.d.*) random variables with zero mean ($E[\epsilon_j] = 0$), representing measurement error as well as other phenomena that cause the observations to deviate from the model predictions $z(t_j;\theta_0)$. The *i.i.d.* assumption means that the errors are uncorrelated across time and have identical variance. We assume the variance is finite and write $\text{var}(\epsilon_j) = \sigma_0^2 < \infty$. We make no further assumptions about the distribution of the errors: specifically, we *do not* assume that they are normally distributed. Under these assumptions, the observation mean is equal to the model prediction, $E[Y_j] = z(t_j;\theta_0)$, while the variance in the observations is a function of the time point, with $\text{var}(Y_j) = z(t_j;\theta_0)^{2\rho}\sigma_0^2$. In particular, this variance is longitudinally *nonconstant* and *model-dependent*. One situation in which this error structure may be appropriate is when observation errors scale with the size of the measurement (so-called *relative noise*), a reasonable scenario in a "counting" process.

Given a set of observations $Y = (Y_1, \ldots, Y_n)$, the estimator $\theta_{GLS} = \theta_{GLS}(Y)$ is defined as the solution of the normal equations

$$\sum_{j=1}^{n} w_j \left[Y_j - z(t_j;\theta)\right] \nabla_\theta z(t_j;\theta) = 0, \tag{7}$$

where the w_j are a set of nonnegative weights [18], defined as

$$w_j = \frac{1}{z(t_j;\theta)^{2\rho}}. \tag{8}$$

The definition in equation (7) assigns different levels of influence, described by the weights, to the different longitudinal observations. Assuming $\rho = 1$ in the error structure described above by Equation (5), we have that the weights are taken to be inversely proportional to the square of the predicted incidence: $w_j = 1/[z(t_j;\theta)]^2$. On the other hand, if $\rho = 1/2$, then the weights are proportional to the reciprocal of the predicted incidence; these correspond to assuming that the variance in the observations is proportional to the value of the model (as opposed to its square). The most popular assumption, the $\rho = 0$ case, leads to the standard ordinary least squares (OLS) approach; see [3] for a full discussion of OLS methods. For the problem and data set we investigate here, the OLS did not produce very reasonable results [15].

Suppose $\{y_j\}_{j=1}^n$ is a realization of the case counting process $\{Y_j\}_{j=1}^n$ and define the function $L(\theta)$ as

$$L(\theta) = \sum_{j=1}^{n} w_j \left[y_j - z(t_j;\theta)\right]^2 . \tag{9}$$

The quantity θ_{GLS} is a random variable, and a realization of it, denoted by $\hat{\theta}_{GLS}$, is obtained by solving

$$\sum_{j=1}^{n} w_j \left[y_j - z(t_j;\theta)\right] \nabla_\theta z(t_j;\theta) = 0, \tag{10}$$

which is *not* equivalent to $\nabla_\theta L(\theta) = 0$ if w_j is given by equation (8) with $\rho \neq 0$; see [3] for further discussion.

Because θ_0 and σ_0^2 are unknown, the estimate $\hat{\theta}_{GLS}$ is used to calculate approximations of σ_0^2 and the covariance matrix Σ_0^n by

$$\sigma_0^2 \approx \hat{\sigma}_{GLS}^2 = \frac{1}{n-4} L(\hat{\theta}_{GLS}) \tag{11}$$

$$\Sigma_0^n \approx \hat{\Sigma}_{GLS}^n = \hat{\sigma}_{GLS}^2 \left[\chi(\hat{\theta}_{GLS},n)^T W(\hat{\theta}_{GLS}) \chi(\hat{\theta}_{GLS},n)\right]^{-1} . \tag{12}$$

In the limit as $n \to \infty$, the GLS estimator has the asymptotic property $\theta_{GLS} \approx \theta_{GLS}^n \sim \mathcal{N}_4(\theta_0, \Sigma_0^n)$ (for details see [3, 18, 34]). Here,

$$W(\hat{\theta}_{GLS}) = \mathrm{diag}(w_1(\hat{\theta}_{GLS}), \ldots, w_n(\hat{\theta}_{GLS})),$$

with $w_j(\hat{\theta}_{GLS}) = 1/[z(t_j;\hat{\theta}_{GLS})]^{2\rho}$. The sensitivity matrix $\chi(\hat{\theta}_{GLS},n)$ denotes the variation of the model output with respect to the parameter, and can be obtained using standard theory [2, 3, 17, 21, 25, 27, 33]. The entries of the j-th row of $\chi(\hat{\theta}_{GLS},n)$ denote how the weekly incidence at time t_j changes in response to changes in the parameter. For example, the first entry of the j-th row of $\chi(\hat{\theta}_{GLS},n)$ is given by (the reader may find further details about the calculation of $\chi(\hat{\theta}_{GLS},n)$ in [15]):

$$\frac{\partial z}{\partial S_0}(t_j;\theta) \quad = \quad \tilde{\beta} \int_{t_{j-1}}^{t_j} \left[I(t;\theta) \frac{\partial S}{\partial S_0}(t;\theta) + S(t;\theta) \frac{\partial I}{\partial S_0}(t;\theta) \right] dt, \tag{13}$$

with $\theta = \hat{\theta}_{GLS}$.

The standard errors for $\hat{\theta}_{GLS}$ can be approximated by taking the square roots of the diagonal elements of the covariance matrix $\hat{\Sigma}_{GLS}^n$.

The values of the weights involved in the GLS estimation depend on the values of the fitted model. These values are not known before carrying out the estimation procedure and consequently the GLS estimation is implemented as an iterative process. The first iteration is carried out by setting $\rho = 0$, which reduces the statistical model in equation (5) to $Y_j = z(t_j;\theta_0) + \epsilon_j$, and also implies the weights in equation (7) are equal to one ($w_j = 1$). This results in an ordinary least squares scheme, the solution of which provides an initial set of weights via equation (8). A weighted least squares fit is then performed using these weights, obtaining updated model values and hence an updated set of weights. The weighted least squares process is repeated until some convergence criterion is satisfied, such as successive values of the estimates being deemed to be sufficiently close to each other. The process can be summarized as follows:

1. Estimate $\hat{\theta}_{GLS}$ by $\hat{\theta}^{(0)}$ using an OLS criterion. Set $k = 0$. Set $\rho = 1$ or $\rho = 1/2$;

2. form the weights $\hat{w}_j = 1/[z(t_j;\hat{\theta}^{(k)})]^{2\rho}$;
3. define $L(\theta) = \sum_{j=1}^{n} \hat{w}_j [y_j - z(t_j;\theta)]^2$. Re-estimate $\hat{\theta}_{GLS}$ by solving
$$\hat{\theta}^{(k+1)} = \arg\min_{\theta \in \Theta} L(\theta)$$
to obtain the $k+1$ estimate $\hat{\theta}^{(k+1)}$ for $\hat{\theta}_{GLS}$;
4. set $k = k+1$ and return to 2. Terminate the procedure when successive estimates for $\hat{\theta}_{GLS}$ are sufficiently close to each other.

The convergence of this procedure is discussed in [9, 18]. This procedure was implemented using a direct search method, the Nelder-Mead simplex algorithm, as discussed by [28], provided by the MATLAB (The Mathworks, Inc.) routine `fminsearch`.

4.2. **Estimation of the effective reproductive number.** Let the pair $(\hat{\theta}, \hat{\Sigma})$ denote the parameter estimate and covariance matrix obtained with the GLS methodology from a given realization $\{y_j\}_{j=1}^{n}$ of the case-counting process. Simulation of the SIR model then allows the time course of the susceptible population, $S(t;\hat{\theta})$, to be generated. The time course of the effective reproductive number can then be calculated as $\mathcal{R}(t;\hat{\theta}) = S(t;\hat{\theta})\hat{\tilde{\beta}}/\hat{\gamma}$. This trajectory is our central estimate of $\mathcal{R}(t)$.

The uncertainty in the resulting estimate of $\mathcal{R}(t)$ can be assessed by repeated sampling of parameter vectors from the corresponding sampling distribution obtained from the asymptotic theory, and applying the above methodology to calculate the $\mathcal{R}(t)$ trajectory that results each time. To generate m such sample trajectories, we sample m parameter vectors, $\theta^{(k)}$, from the 4-multivariate normal distribution $\mathcal{N}_4(\hat{\theta}, \hat{\Sigma})$. We require that each $\theta^{(k)}$ lies within a feasible region Θ determined by biological constraints. If this is not the case for a particular sample, we discard it and then we resample until $\theta^{(k)} \in \Theta$. Numerical solution of the SIR model using $\theta^{(k)}$ allows the sample trajectory $\mathcal{R}(t;\theta^{(k)})$ to be calculated. We summarize these steps involved in the construction of the sampling distribution of the effective reproductive number:

1. Set $k = 1$;
2. obtain the k-th parameter sample from the 4-multivariate normal distribution:
$$\theta^{(k)} \sim \mathcal{N}_4(\hat{\theta}, \hat{\Sigma});$$
3. if $\theta^{(k)} \notin \Theta$ (constraints are not satisfied) return to 2. Otherwise go to 4;
4. using $\theta = \theta^{(k)}$ find numerical solutions, denoted by $\big(S(t;\theta^{(k)}), I(t;\theta^{(k)})\big)$, to the nonlinear system defined by Equations (1) and (2). Construct the effective reproductive number as follows:
$$\mathcal{R}(t;\theta^{(k)}) = S(t;\theta^{(k)})\frac{\tilde{\beta}^{(k)}}{\gamma^{(k)}},$$
where $\theta^{(k)} = \Big(S_0^{(k)}, I_0^{(k)}, \tilde{\beta}^{(k)}, \gamma^{(k)}\Big)$;
5. set $k = k+1$. If $k > m$ then terminate. Otherwise return to 2.

Uncertainty estimates for $\mathcal{R}(t)$ are calculated by finding appropriate percentiles of the distribution of the $\mathcal{R}(t)$ samples.

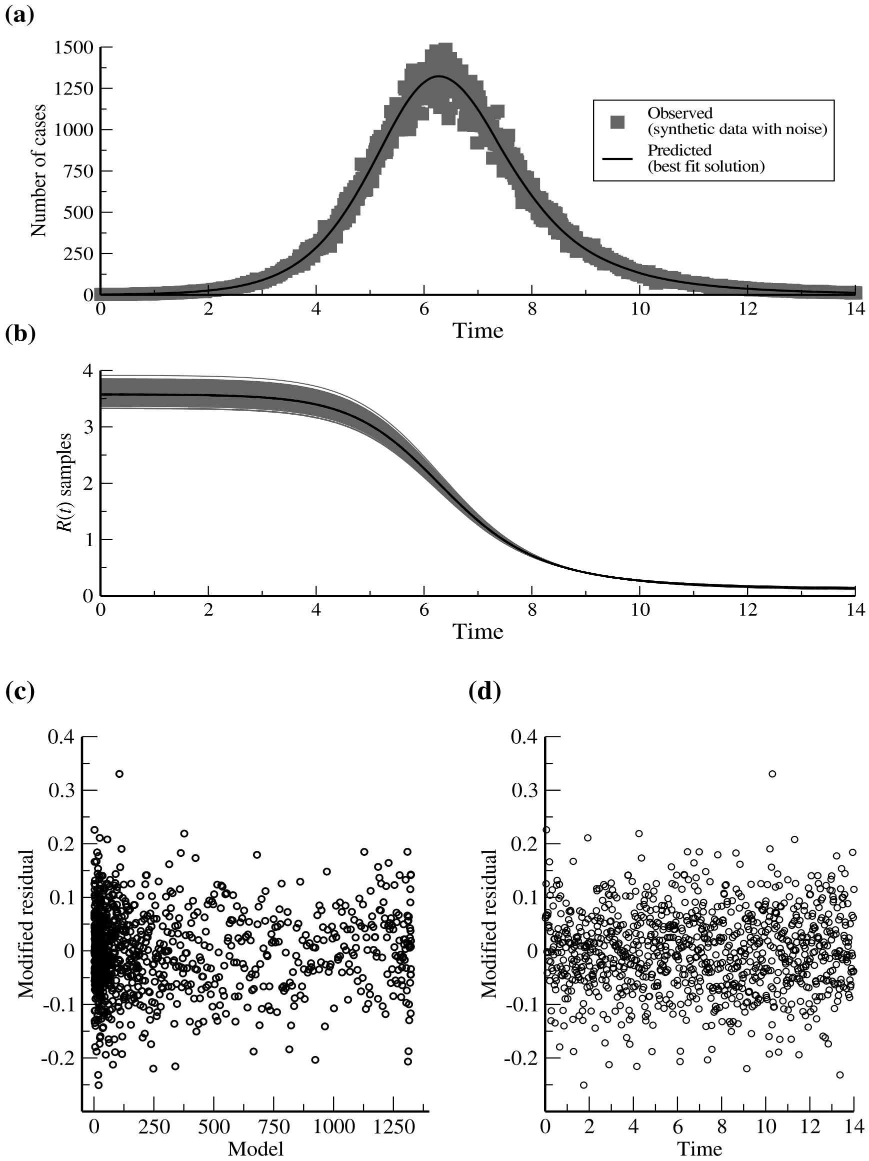

FIGURE 2. Results from applying the GLS methodology to synthetic data with non-constant variance noise ($\alpha = 0.075$), using $n = 1,000$ observations. The initial guess for the optimization routine was $\theta = 1.10\theta_0$. The weights in the cost function were equal to $1/z(t_j;\theta)^2$, for $j = 1, \ldots, n$. Panel (a) depicts the observed and fitted values and panel (b) displays $1,000$ of the $m = 10,000$ $\mathcal{R}(t)$ sample trajectories. Residuals plots are presented in panels (c) and (d): modified residuals versus fitted values in (c) and modified residuals versus time in (d).

TABLE 2. Estimates from a synthetic data set of size $n = 1,000$, with non-constant variance using $\alpha = 0.075$. The $\mathcal{R}(t)$ sample size is $m = 10,000$. The initial guess of the optimization algorithm was $\theta = 1.10\theta_0$. Each weight in the cost function $L(\theta)$ (see Equation (9)) was equal to $1/z(t_j;\theta)^2$ for $j = 1,\ldots,n$. The units of the estimated quantities are: people, for S_0 and I_0; per person per week, for $\tilde{\beta}$; and per week, for γ.

Parameter	True value	Initial guess	Estimate	Standard error
S_0	3.500×10^5	3.800×10^5	3.498×10^5	1.375×10^3
I_0	9.000×10^1	9.900×10^1	9.085×10^1	1.424×10^0
$\tilde{\beta}$	5.000×10^{-6}	5.500×10^{-6}	4.954×10^{-6}	4.411×10^{-8}
γ	5.000×10^{-1}	5.500×10^{-1}	4.847×10^{-1}	1.636×10^{-2}
		$L(\hat{\theta}_{GLS}) = 5.689\times10^0$		
$\sigma_0^2 = 5.625\times10^{-3}$				$\hat{\sigma}^2_{GLS} = 5.712\times10^{-3}$
	Min.$\mathcal{R}(t;\hat{\theta}_{GLS})$		0.132 [0.120,0.146]	
	Max.$\mathcal{R}(t;\hat{\theta}_{GLS})$		3.576 [3.420,3.753]	
True value of the reproductive number at time t_0; $\mathcal{R}(t_0) = S_0\tilde{\beta}/\gamma = 3.500$				

5. **Estimation scheme applied to synthetic data.** We generated a synthetic data set with nonconstant variance noise. The true value θ_0 was fixed, and was used to calculate the numerical solution $z(t_j;\theta_0)$. Observations were computed in the following fashion:

$$Y_j = z(t_j;\theta_0) + z(t_j;\theta_0)\alpha V_j = z(t_j;\theta_0)\left(1+\alpha V_j\right), \tag{14}$$

where the V_j are independent random variables with standard normal distribution (i.e., $V_j \sim \mathcal{N}(0,1)$), and $0 < \alpha < 1$ denotes a desired percentage. Hence $\rho = 1$ in the general formulation with $\epsilon_j = \alpha V_j$. In this way, $\mathrm{var}(Y_j) = [z(t_j;\theta_0)\alpha]^2$ which is nonconstant across the time points t_j. If the terms $\{v_j\}_{j=1}^n$ denote a realization of $\{V_j\}_{j=1}^n$, then a realization of the observation process is denoted by $y_j = z(t_j;\theta_0)(1+\alpha v_j)$.

An $n = 1,000$ point synthetic data set was constructed with $\alpha = 0.075$. The optimization algorithm was initialized with the estimate $\theta = 1.10\theta_0$. The weights in the normal equations defined by Equation (7), were chosen as $w_j = 1/z(t_j;\theta)^2$ (i.e., $\rho = 1$).

Table 2 lists estimates of the parameters and $\mathcal{R}(t)$, together with uncertainty estimates. In the case of $\mathcal{R}(t)$, uncertainty was assessed based on the simulation approach using $m = 10,000$ samples of the parameter vector, drawn from $\mathcal{N}_4(\hat{\theta}_{GLS},\hat{\Sigma}^n_{GLS})$. Fig. 2(a) depicts both data and fitted model points $z(t_j;\hat{\theta}_{GLS})$ plotted versus t_j. Fig. 2(b) depicts 1,000 of the 10,000 $\mathcal{R}(t)$ curves.

Residuals plots are displayed in Fig. 2(c) and (d). Because $\alpha v_j = (y_j - z(t_j;\theta_0))/z(t_j;\theta_0)$, by construction of the synthetic data, the residuals analysis focuses on the ratios

$$\frac{y_j - z(t_j;\hat{\theta}_{GLS})}{z(t_j;\hat{\theta}_{GLS})},$$

which in the labels of Fig. 2(c) and (d) are referred to as "Modified residuals" (for a more detailed discussion of residuals and modified residuals, see [3]). In Fig. 2(c) these ratios are plotted against $z(t_j;\hat{\theta}_{GLS})$, while Panel (d) displays them versus the time points t_j. The lack of any discernable patterns or trends in Fig. 2(c) and (d) suggests that the errors in the synthetic data set conform to the assumptions made in the formulation of the statistical model of equation (14). In particular, the errors are uncorrelated and have variance that scales according to the relationship stated above.

6. **Analysis of influenza outbreak data.** The GLS methodology was applied to longitudinal observations of six influenza outbreaks (see Section 2), giving estimates of the parameters and the reproductive number for each season. The number of observations n varies from season to season. The $\mathcal{R}(t)$ sample size was $m = 10,000$ in each case. The set of admissible parameters Θ is defined by the lower and upper bounds listed in Table 3 along with the inequality constraint $S_0\tilde{\beta}/\gamma > 1$. The bounds in Table 3 were obtained from or based on [10, 29, 32] and references therein. For brevity, we only present here the results obtained using data from the 1998–99 season with GLS methods. Further results including (unsuccessful) use of OLS methodology can be found in [15].

TABLE 3. Lower and upper bounds on the initial conditions and parameters.

Suitable range	Unit
$1.00\times10^2 < S_0 < 7.00\times10^6$	people
$0.00 < I_0 < 5.00\times10^3$	people
$7.00\times10^{-9} < \tilde{\beta} < 7.00\times10^{-1}$	weeks^{-1}people^{-1}
$3/7 < 1/\gamma < 4/7$	weeks

Visual inspection suggests that the model fits obtained using the GLS approach (Fig. 3) are even worse than those obtained using OLS (the results obtained using OLS can be found in [15]). This is somewhat misleading, however, because the weights, defined as $w_j = 1/[z(t_j;\theta)]^2$, mean that the GLS fitting procedure (unlike visual inspection of the figures) places increased emphasis on datapoints whose model value is small and decreased emphasis on datapoints where the model value is large. If these graphs are, instead, plotted with a logarithmic scale on the vertical axis, an accurate visualization is obtained (Fig. 4): multiplicative observation

TABLE 4. Results of GLS estimation applied to influenza data from season 1998–99, weights equal to $1/z(t_j;\theta)^2$.

Parameter	Estimate	Unit	Standard error
S_0	7.939×10^3	people	1.521×10^4
I_0	2.436×10^{-1}	people	4.216×10^{-1}
$\tilde{\beta}$	3.458×10^{-4}	weeks^{-1}people	5.233×10^{-5}
γ	2.333×10^0	weeks^{-1}	5.318×10^0
$L(\hat{\theta}_{GLS}) = 1.754\times10^1$ $\hat{\sigma}^2_{GLS} = 6.047\times10^{-1}$			
Min.$\mathcal{R}(t;\hat{\theta}_{GLS})$		0.843	[0.784,1.018]
Max.$\mathcal{R}(t;\hat{\theta}_{GLS})$		1.177	[1.052,1.252]

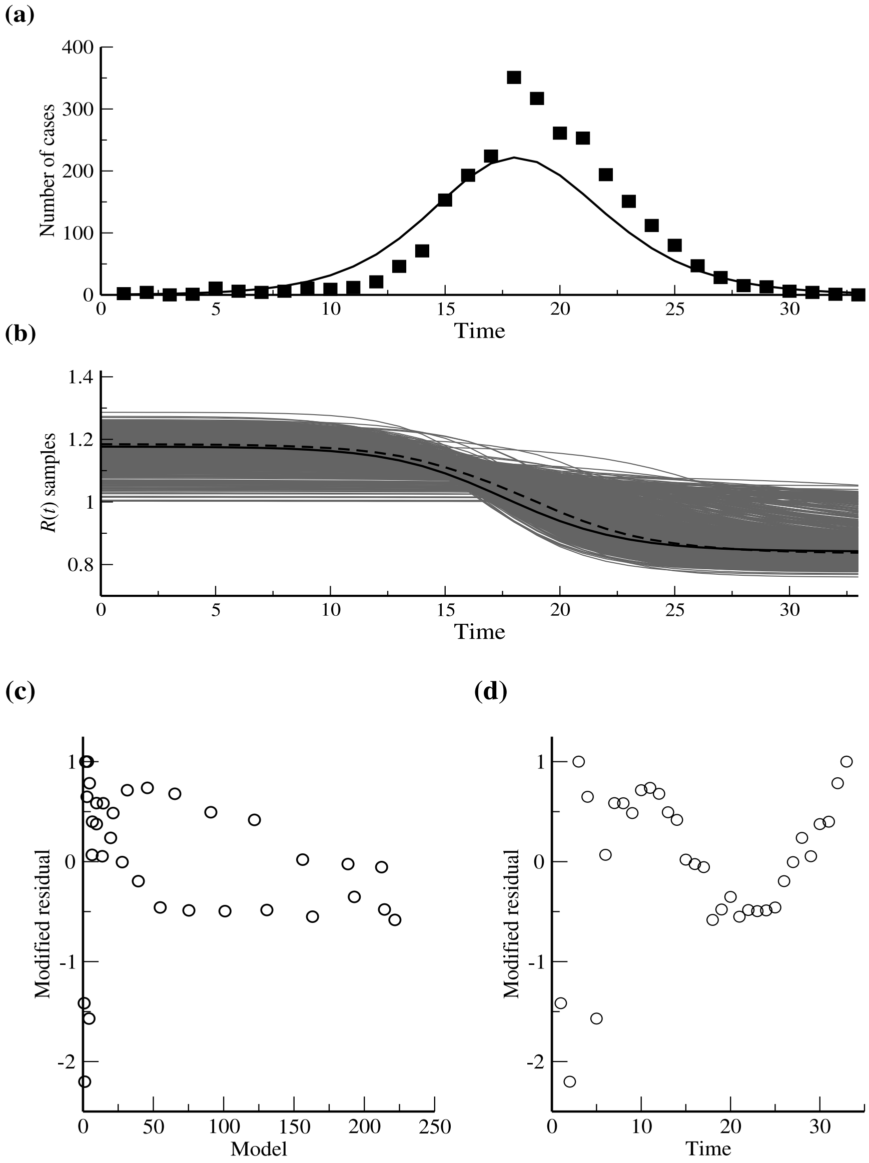

FIGURE 3. GLS applied to influenza data from 1998–99 season. The weights were taken equal to $1/z(t_j;\theta)^2$. Panel (a) depicts the observations (solid squares) as well as the model prediction (solid curve). In Panel (b) $1,000$ of the $m = 10,000$ samples of the effective reproductive number $\mathcal{R}(t)$ are displayed. The solid curve depicts the central estimate $\mathcal{R}(t;\hat{\theta}_{GLS})$ and the dashed curve the median of the $\mathcal{R}(t)$ samples at each point in time. Panel (c) exhibits the modified residuals $(y_j - z(t_j;\hat{\theta}_{GLS}))/z(t_j;\hat{\theta}_{GLS})$ plotted versus the model predictions, $z(t_j;\hat{\theta}_{GLS})$. Panel (d) displays the modified residuals plotted against time.

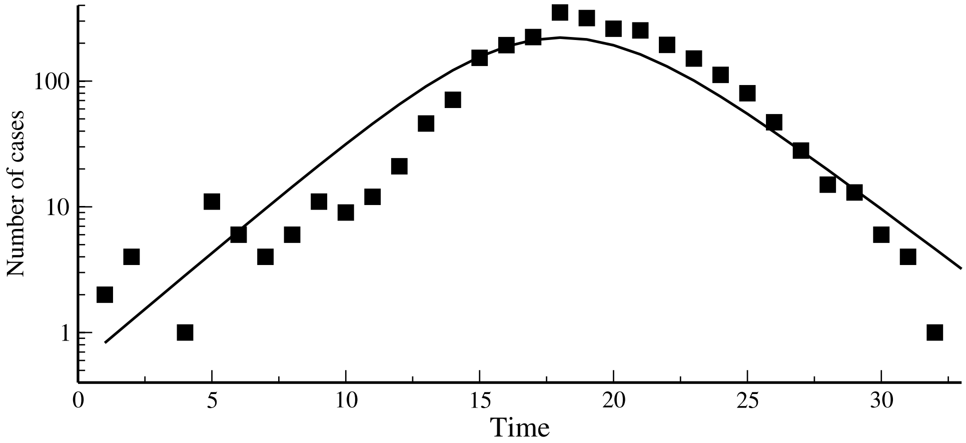

FIGURE 4. Best-fitting model for the 1998–99 season, obtained using GLS with $1/z(t_j;\theta)^2$ weights. Observations (solid squares) and the model prediction (solid curve) are plotted on a logarithmic scale.

noise on a linear scale becomes constant variance additive observation noise on a logarithmic scale.

The parameter estimates have standard errors that are often of the same order of magnitude as the estimates themselves (Table 4). The residuals plots reveal clear patterns and trends (Fig. 3(c) and (d)). Temporal trends in the residuals (and visual inspection of the plots depicting the best fitting model and the datapoints) indicate that there are systematic differences between the fitted model and the data. For instance, it appears that the fitted model peaks slightly earlier than the observed outbreak, and, as a result, there are numbers of sequential points where the data lies above or below the model. The modified residuals versus model plot suggests that the variation of the residuals may be decreasing as the model value increases.

The condition number of the matrix $\chi(\hat{\theta}_{GLS},n)^T W(\hat{\theta}_{GLS})\chi(\hat{\theta}_{GLS},n)$ is 9.0×10^{19}. This is very similar to that for the OLS estimation (shown in [15]), again suggesting caution in interpreting the standard errors.

It is quite plausible that our description of the error structure of the data is inadequate when the numbers of cases are at low levels (a not uncommon situation), so that the *statistical model* chosen is not correct. In particular, the reporting process might change as the outbreak emerges (e.g., doctors become more alert to possible flu cases) or comes close to ending. Moreover, our *mathematical model* may also be incorrect because it is deterministic whereas an epidemic contains stochasticity. Stochastic effects may exhibit a relatively large impact at the beginning or end of an epidemic, when the numbers of cases are low. It is possible for the infection to undergo extinction, a phenomenon which cannot be captured by the deterministic model. Spatial clustering of cases is also a distinct possibility, particularly during the early stages of an outbreak. This will affect the time course of an outbreak as well as the reporting process: clustering of cases may well increase the reporting noise if cases in a cluster tend to get reported together (e.g., a cluster occurs within

TABLE 5. Estimation results from GLS, with weights $1/z(t_j;\theta)$, applied to truncated influenza data set for season 1998–99.

Parameter	Estimate	Unit	Standard error
S_0	6.017×10^3	people	3.287×10^3
I_0	2.091×10^0	people	9.483×10^{-1}
$\tilde{\beta}$	3.797×10^{-4}	weeks^{-1}people	1.774×10^{-5}
γ	1.750×10^0	weeks^{-1}	1.317×10^0
$L(\hat{\theta}_{GLS}) = 3.872\times10^1$			
$\hat{\sigma}^2_{GLS} = 2.277\times10^0$			
Min.$\mathcal{R}(t;\hat{\theta}_{GLS})$		0.750	[0.748,0.819]
Max. $\mathcal{R}(t;\hat{\theta}_{GLS})$		1.306	[1.212,1.308]

an area where many isolates are sent to the CDC) or not reported together (e.g., a cluster occurs in an area that has poorer coverage in the reporting process).

Indeed, examination of one of the influenza time series plotted on a logarithmic scale (Fig. 4) indicates that both the beginning and end of the time series are problematic. The fit of the model (see [15] for additional details) is clearly poorer over these parts of the time series, which correspond to the times when the observed values are small.

Both forms of the weights (inversely proportional to the square of the predicted incidence or inversely proportional to the predicted incidence) mean that errors at these small values have considerable impact on the cost function, and hence on the GLS estimation process, although this is less of a concern for the $1/z$ weights.

Another issue that has been raised by studies of parameter estimation in biological situations concerns redundancy in information measured when a system is close to its equilibrium [4]. This might be a relevant issue for the final part of the outbreak data, as there is often a period lasting ten or more weeks when there are few cases.

We investigated whether the removal of the lowest-valued points from the data sets would improve the inverse problem results. We constructed truncated data sets by considering only the period between the time when the number of isolates first reached ten at the beginning of the outbreak and first fell below ten at the end of the outbreak. As a notational convenience, we refer to the numbers of susceptibles and infectives at the start of the first week of the truncated data set as S_0 and I_0, even though these times no longer correspond to the start of the influenza season. (For example, in Fig. 5, S_0 and I_0 refer to the state of the system at $t = 8$.)

Using fewer observations, with the $1/z$ weights, we obtained a decrease in the standard errors for most of the parameter estimates (comparing Tables 4 and 5). This decrease occurs even though the number of points in the data set has fallen from 35 to 23, causing the factor $1/(n-4)$ that appears in equation (11) to increase by 80%. The corresponding residuals plots (see Fig. 5(c) and (d)) provide no suggestion that the assumptions of the statistical model are invalid (contrasting Fig. 3(c) and (d), which display temporal trends), and hence we conclude the statistical model with $\rho = 1/2$ might be reasonable.

The condition number of the matrix $\chi(\hat{\theta}_{GLS}, n)^T W(\hat{\theta}_{GLS})\chi(\hat{\theta}_{GLS}, n)$ is 9.2×10^{19}. Truncation of the data sets helped considerably with the GLS estimation process,

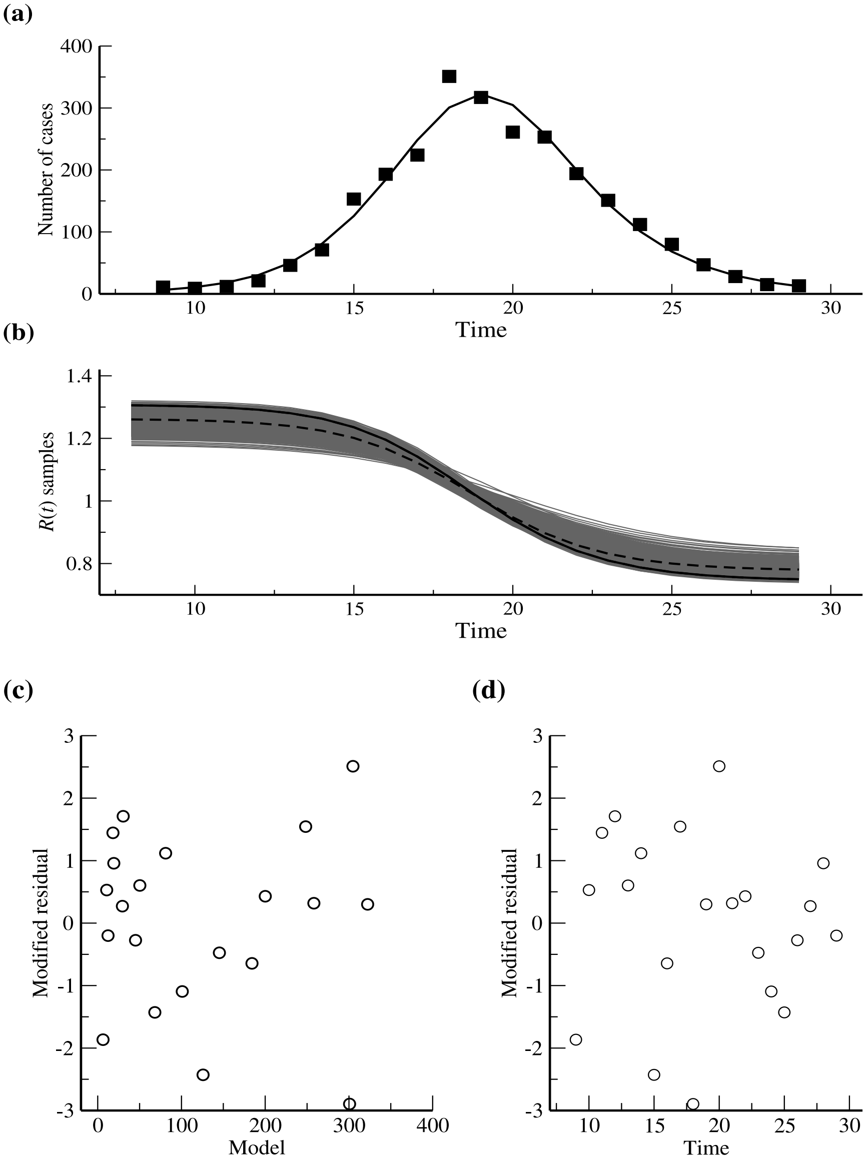

FIGURE 5. Model fits obtained using GLS on truncated influenza data from season 1998–99, weights equal to $1/z(t_j;\theta)$. Panel (a) shows the observations (solid squares) as well as the model prediction (solid curve). In Panel (b) $1,000$ of the $m = 10,000$ samples of the effective reproductive number $\mathcal{R}(t)$ are displayed together with the central estimate $\mathcal{R}(t;\hat{\theta}_{GLS})$ (solid curve) and the median of the $\mathcal{R}(t)$ samples at each time point (dashed curve). Panel (c) shows the modified residuals $(y_j - z(t_j;\hat{\theta}_{GLS}))/z(t_j;\hat{\theta}_{GLS})^{1/2}$ versus the model prediction. In panel (d), each modified residual is displayed versus the observation time point.

although the large condition numbers might be cause for caution in relying too much on the standard errors.

Truncation of the data set had little effect on the parameter estimates obtained using OLS (results are detailed in [15]), except that the values of S_0 and I_0 were changed because they refer to a later initial time, as discussed above. Standard errors for the OLS estimates were higher with the truncated data set than for the full data set, as should be expected given the reduced number of data points. Overall, the results using OLS with the truncated data were less than satisfactory.

7. **Discussion.** We have discussed parameter estimation methodologies (OLS, GLS with different weighting factors) that, using sensitivity analysis and asymptotic statistical theory, also provide measures of uncertainty for the estimated parameters. The GLS techniques were illustrated first using synthetic data sets, and it was seen that they can perform very well with reasonable data sets. Even within the ideal situation provided by synthetic data, potential problems of the approach were identified [15]. Worryingly, these problems were not apparent from inspection of the uncertainty estimates (standard errors) alone. However, these problems were revealed by examination of model fit diagnostic plots, constructed in terms of the residuals of the fitted model (see [15]). Using these post-analysis residual plots, we were able to identify a likely statistical error model for the particular influenza surveillance data set. The results here and in [15] argue strongly for the routine use of uncertainty estimation, together with careful examination of residuals plots when using SIR-type models with surveillance data.

Development of mathematical and statistical methods geared to the (real-time) estimation of the effective reproductive number is relatively widespread and growing, with various contributions including [5, 6, 11, 24, 31, 39]. Wallinga and Teunis [39] developed a likelihood-based methodology that assumes the generation interval (time from symptom onset in a primary case to symptom onset in a secondary case) is described by a Weibull distribution and that a specific infection network underlies the observed epidemic curve; a likelihood-based procedure infers who infected whom, from pairs of cases rather than the entire infection network. Cauchemez, et al., [11] proposed a methodology to monitor the efficacy of outbreak intervention; an outbreak is under control if estimates of the effective reproductive number $\mathcal{R}(t)$ are below unity; $\mathcal{R}(t) < 1$. The proposed method involves: data on the number of cases (incidence), data on the time of onset of symptoms, and contact tracing information from a subset of cases. A Markov chain Monte Carlo (MCMC) scheme is used to estimate the posterior distribution of the generation interval, and eventually $\mathcal{R}(t)$, up to the last observation. Real-time $\mathcal{R}(t)$ estimates can only be calculated after having the estimated posterior distribution of the generation interval. Forsberg White and Pagano [24] devised a method that uses simple surveillance data to estimate the basic reproductive number and the serial interval (also referred to as the generation interval). This methodology is likelihood-based; the likelihood of the observed counts of cases of infection is based on an evolving Poisson distribution, from which the maximum likelihood estimates (MLE) are derived. Additionally, branching process theory is used to calculate an estimator that is contrasted to the MLE and to the Bayesian posterior mode (with informative and noninformative priors); all of these are illustrated using simulated observations. The simultaneous estimation of the serial interval (estimated along with the basic reproductive number) does not require information about contact tracing. However, it is assumed that the distribution of the serial interval is gamma; the methodology can be adjusted to model the serial interval with a different parametric model. Nishura [31] estimated the effective reproductive number by applying a discrete-time branching process to back-calculated incidence data, assuming three different serial intervals. The absence of temporal monotonic decrease in the reproductive number estimates is suggestive of time variation in the patterns of secondary transmission. Bettencourt, et al., [5] and Bettencourt and Ribeiro [6] formulated stochastic models for the time evolution of the number of cases in the context of emerging diseases. In these formulations the effective reproductive number is a time-evolving parameter which is inferred by calculating its posterior distribution (application of a Bayesian scheme) at each observation time point using the observed number of cases at hand. Their proposed methodology addresses uncertainty quantification of prediction along with anomaly detection (two-sided p-value significance test).

The statistical methodology presented here addresses the effect of observation error on parameter estimation. While the approach can handle different statistical models for the observation process, it does assume that we have a mathematical model that correctly describes the behavior of the system, albeit for an unknown value of the parameter vector. The methodology does not examine the effect of mis-specification of the mathematical model. It is well known that this effect can often dwarf the uncertainty that arises from observation error [30]. Examination of residual plots, however, can identify systematic deviations between the behavior of the model and the data.

The methodology proposed here is based on a deterministic formulation (single-outbreak SIR) of the underlying epidemic process. Consequently, the constructed curves of the effective reproductive number are deterministic in nature (while the uncertainty quantification results from the statistical model of the observation error), and as such they show monotonically decreasing temporal patterns; from the beginning to the end of an outbreak. It is clear that our methodology would fail to reflect any stochastic behavior in $\mathcal{R}(t)$. In fact, incidence curves exhibiting bimodality and strong stochasticity are most likely not suitable for the application of our methodology as it stands, unless either the rescaled transmission rate, $\tilde{\beta}$, or the recovery rate, γ, or both, are modeled as time-dependent coefficients, that is, unless we modify the mathematical model.

Another limitation of the modeling methodology proposed here is that we use an SIR model which neglects a latency period (a delay prior to the development of active infection; a stage where individuals become capable of transmitting infection to others). It has been suggested before [14, 30] that ignoring the latency period may result in biased estimations of the reproductive number (an illustration of the effect of a latency period on $\mathcal{R}(t)$ is given in the appendix). The estimation framework presented here could be readily applied to mathematical models with latency.

Application of several least squares approaches to the influenza isolate data gave mixed results (applications of both OLS and GLS are addressed more fully in [15]). Estimates of the effective reproductive number were in broad agreement with results obtained in other studies (see Table 6). While apparently reasonable fits were obtained in some instances, the uncertainty analyses highlighted situations in which visual inspection suggested that a good fit had been obtained but for which estimated parameters had large uncertainties. Residual plots showed that variance in the surveillance data may not have been constant (i.e., observation noise was not simply additive, $\mathrm{var}(Y_j) \neq \sigma_0^2$), but more likely scaled according to either the square

Table 6. Comparison between reproductive number estimates across studies of interpandemic influenza. In this table $\mathcal{R}_0$ stands for the basic reproductive number (naïve population), while $\max(\mathcal{R}(t))$ denotes the initial effective reproductive number in a non-naïve population.

Studies of interpandemic influenza	Estimates
Bonabeau et al. [7]	$1.70 \leq \mathcal{R}_0 \leq 3.00$
Chowell et al. [13]	$1.30 \leq \max(\mathcal{R}(t)) \leq 1.50$
Dushoff et al. [20]	$4.00 \leq \mathcal{R}_0 \leq 16.00$
Flahault et al. [23]	$\mathcal{R}_0 = 1.37$
Spicer & Lawrence [35]	$1.46 \leq \mathcal{R}_0 \leq 4.48$
Viboud et al. [38]	$1.90 \leq \max(\mathcal{R}(t)) \leq 2.50$

of the fitted model value (i.e., relative measurement error, $\text{var}(Y_j) = z(t_j;\theta_0)^2\sigma_0^2$) or the fitted model value itself (i.e., $\text{var}(Y_j) = z(t_j;\theta_0)\sigma_0^2$). The potentially large impact of errors at low numbers of cases on the GLS estimation process was clearly observed.

Temporal trends were observed in some of the residuals plots, indicative of systematic differences between the behavior of the SIR model and the data. Potential sources of these differences include inadequacies of the mathematical model to describe the process underlying the data and issues with the reliability of (i.e., variability in) the data itself. We emphasize, however, that our use of these typical data sets provide an excellent *illustration* of the methodologies as well as the possible pitfalls that may be inherent in attempting to use typical surveillance data to estimate parameters and effective reproductive numbers.

Sophisticated mathematical and statistical algorithms and analyses can be utilized to fit SIR-type epidemiological models to surveillance data. Reasonable quality data, good mathematical and statistical models, and careful post analyses using residual plots are all required if this approach is to be successful. In many instances, however, the available surveillance data is most likely inadequate to validate the SIR model with any degree of confidence especially if a mildly inadequate mathematical model and an incorrect statistical model for the data are chosen. This is likely to be true in much of the modeling efforts reported for epidemics where the data collection process has inadequacies and where no uncertainty quantification along with post analysis are done.

Acknowledgements. This material was based upon work supported in part by the Statistical and Applied Mathematical Sciences Institute (via a SAMSI postdoctoral fellowship to A. C.-A.) which is funded by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This research was also supported in part by Grant Number R01AI071915-07 from the National Institute of Allergy and Infectious Diseases. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or the NIH. The authors are thankful for the opportunity to contribute in this special edition honoring Karl Hadeler and Fred Brauer.

Appendix. It is well known that approaches based on model fitting lead to underestimates of the basic reproductive number of an infection if the latent period of the infection is ignored, *i.e.*, if an SIR model is used to describe an outbreak when an SEIR model would have been more appropriate (see [14, 30, 40] and references therein).

We illustrate the effects of infection latency on the estimation of $\mathcal{R}(t)$ by considering a synthetic data set obtained using the standard single outbreak SEIR model

$$\frac{dS}{dt} = -\tilde{\beta}SI \tag{15}$$

$$\frac{dE}{dt} = \tilde{\beta}SI - \alpha E \tag{16}$$

$$\frac{dI}{dt} = \alpha E - \gamma I \tag{17}$$

$$\frac{dX}{dt} = \gamma I. \tag{18}$$

Here, $\tilde{\beta} = \beta/N$ and we take the initial condition to be $(S(t_0) = S_0, E(t_0) = 0, I(t_0) = I_0, X(t_0) = 0)$. The parameter α denotes the rate at which individuals progress from the latent class E to the infectious class I.

The effective reproductive number for this model is given by

$$\mathcal{R}(t) \equiv \mathcal{R}(t;q) = S(t;q)\tilde{\beta}/\gamma, \tag{19}$$

just as for the SIR model, but where $q = (S_0, I_0, \tilde{\beta}, \gamma, \alpha)$.

Latency slows the spread of infection: time spent in the latent class means an individual's secondary infections occur later than they would if there was no latency. If SIR and SEIR models were simulated using the same set of parameters and initial conditions, the outbreak would occur more rapidly for the SIR model. Consequently, if we considered the forward problem and calculated $\mathcal{R}(t)$ curves for corresponding SIR and SEIR models, we would see that their initial values would be identical but that there would be a more rapid decrease in $\mathcal{R}(t)$ for the SIR model as its susceptible population is more rapidly depleted. The situation is not so simple, however, for the inverse problem because parameter values are estimated from the data: we would not expect to obtain the same set of parameter values if we fitted the two different models.

Our synthetic data set was generated by adding constant variance noise to an incidence time series obtained from the SEIR model (for details in calculating synthetic data see [15]). The OLS procedure was used to estimate the parameter vector $q = (S_0, I_0, \tilde{\beta}, \gamma, \alpha)$ for the SEIR model and the vector $\theta = (S_0, I_0, \tilde{\beta}, \gamma)$ for the SIR model.

In Fig. 6 we display the central estimates of $\mathcal{R}(t)$ obtained using the two models: the SIR-based estimates $\mathcal{R}(t;\hat{\theta})$ (circles), and the SEIR-based estimates $\mathcal{R}(t;\hat{q})$ (crosses). Using our known parameter values (listed in the figure caption), the true value of the reproductive number at time t_0 is $\mathcal{R}(t_0) = S_0\tilde{\beta}/\gamma = 2.45$. Use of the SEIR model provides us with a good estimate of this quantity, while the SIR-based approach leads to an appreciable underestimate, in accordance with the well-known results discussed above.

Over the course of the outbreak, both $\mathcal{R}(t;\hat{\theta})$ and $\mathcal{R}(t;\hat{q})$ decrease as the susceptible population becomes depleted. Because its initial value is greater, the SEIR-based

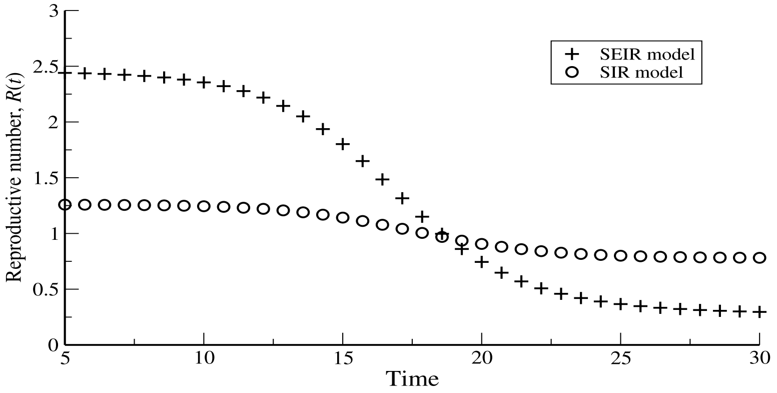

FIGURE 6. OLS estimates, obtained from synthetic data, of the effective reproductive number, $\mathcal{R}(t)$, versus time, t. The circles (single-outbreak SIR model) display $\mathcal{R}(t) \equiv \mathcal{R}(t;\hat{\theta}) = S(t;\hat{\theta})\hat{\tilde{\beta}}/\hat{\gamma}$ with $\hat{\theta} = (\hat{S}_0, \hat{I}_0, \hat{\tilde{\beta}}, \hat{\gamma})$, while the crosses (single-outbreak SEIR model) display $\mathcal{R}(t) \equiv \mathcal{R}(t;\hat{q}) = S(t;\hat{q})\hat{\tilde{\beta}}/\hat{\gamma}$ where $\hat{q} = (\hat{S}_0, \hat{I}_0, \hat{\tilde{\beta}}, \hat{\gamma}, \hat{\alpha})$. The parameter values used to generate the synthetic data were $\tilde{\beta} = 3.5 \times 10^{-6}$, $\gamma = 0.50$, $\alpha = 1.5$, $S_0 = 3.50 \times 10^5$, and $I_0 = 90.0$. The true reproductive number at time t_0 is 2.45.

estimate $\mathcal{R}(t;\hat{q})$ falls by a greater amount than the SIR-based estimate $\mathcal{R}(t;\hat{\theta})$. For both models, the estimated number of susceptibles falls by almost the same amount, which is unsurprising given that the decrease in the number of susceptibles is equal to the total incidence over the outbreak.

Residuals plots give an indication of the inadequacy of the SIR model as a description of the synthetic data set: temporal patterns are clearly visible when the SIR residuals are plotted against time. No such pattern is seen in the corresponding plot of the residuals from the SEIR model fit.

This synthetic data example illustrates that use of an inadequate mathematical description of the epidemic process can be misleading. Because influenza infection has a latent period, an SEIR model is likely to be a more appropriate choice than an SIR model and so the estimates we obtained in the main text should, therefore, be interpreted with some caution. Having said this, the methodological issues that are the main part of this study, namely the statistical uncertainty analysis and the diagnostic information provided by residuals plots, apply regardless of the mathematical model that is employed.

REFERENCES

[1] R. Anderson and R. May, "Infectious Diseases of Humans: Dynamics and Control," Oxford University Press, 1991.

[2] P. Bai, H. T. Banks, S. Dediu, A. Y. Govan, M. Last, A. L. Lloyd, H. K. Nguyen, M. S. Olufsen, G. Rempala and B. D. Slenning, *Stochastic and deterministic models for agricultural production networks*, Math. Biosci. Eng., **4** (2007), 373–402.

[3] 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-TR08-1, NCSU, January, 2008; in "Mathematical and Statistical Estimation Approaches in Epidemiology" (eds. G. Chowell, et. al.), Springer, New York, to appear.

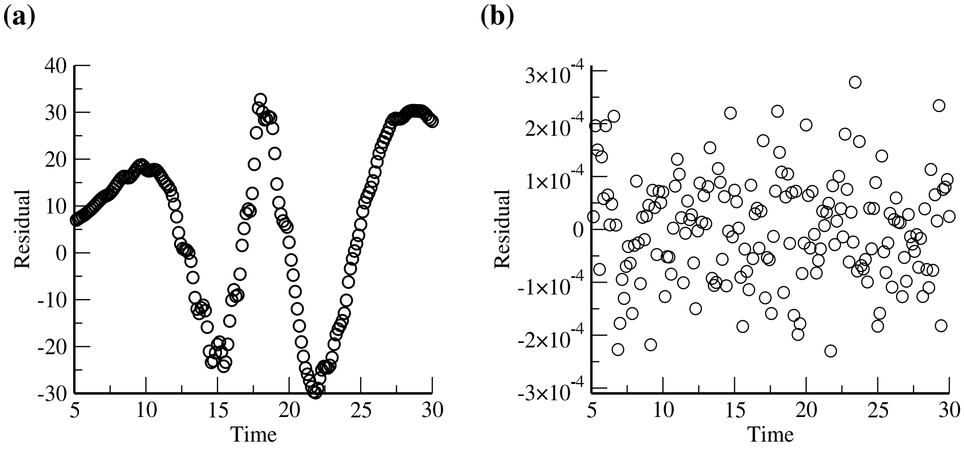

FIGURE 7. Residuals plots from OLS estimation applied to the SEIR-generated synthetic data set. (a) Residuals versus time for the SIR-based estimation. (b) Residuals versus time for the SEIR-based estimation.

[4] H. T. Banks, S. L. Ernstberger and S. L. Grove, *Standard errors and confidence intervals in inverse problems: Sensitivity and associated pitfalls*, J. Inv. Ill-posed Problems, **15** (2007), 1–18.

[5] L. M. A. Bettencourt and R. M. Ribeiro, *Real time Bayesian estimation of the epidemic potential of emerging infectious diseases*, PLOS One, **3** (2008), e2185. DOI: 10.1371/journal.pone.0002185.

[6] L. M. A. Bettencourt, R. M. Ribeiro, G. Chowell, T. Lant and C. Castillo-Chavez, *Towards real time epidemiology: Data assimilation, modeling and anomaly detection of health surveillance data streams*, in "Lecture Notes in Computer Science" (eds. D.Zeng, I. Gotham, K. Komatsu and C. Lynch), Vol. **4506** (2007), 79–90.

[7] E. Bonabeau, L. Toubiana and A. Flahault, *The geographical spread of influenza*, Proc. R. Soc. Lond. B, **265** (1998), 2421–2425.

[8] F. Brauer and C. Castillo-Chávez, "Mathematical Models in Population Biology and Epidemiology," Springer, New York, 2001.

[9] R. J. Carroll, C. F. Wu and D. Ruppert, *The effect of estimating weights in weighted least squares*, J. Am. Stat. Assoc., **83** (1988), 1045–1054.

[10] S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron and P. Y. Boelle, *A Bayesian MCMC approach to study transmission of influenza: Application to household longitudinal data*, Stat. Med., **23** (2004), 3469–3487.

[11] S. Cauchemez, P. Boelle, G. Thomas and A. Valleron, *Estimating in real time the efficacy of measures to control emerging communicable diseases*, Am. J. Epidemiol., **164** (2006), 591–597.

[12] Centers for Disease Control and Prevention (CDC), Flu activity, reports and surveillance methods in the United States, website: http://www.cdc.gov/flu/weekly/fluactivity.htm, accessed on April 7, 2006.

[13] G. Chowell, M. A. Miller and C. Viboud, *Seasonal influenza in the United States, France, and Australia: transmission and prospects for control*, Epidemiol. Infect., **136** (2008), 852–864.

[14] G. Chowell, H. Nishiura and L. M. A. Bettencourt, *Comparative estimation of the reproduction number for pandemic influenza from daily case notification data*, J. Roy. Soc. Interface, **4** (2007), 155–166.

[15] A. Cintrón-Arias, C. Castillo-Chávez, L. M. Bettencourt, A. L. Lloyd and H. T. Banks, *The estimation of the effective reproductive number from disease outbreak data*, Center for Research in Scientific Computation Technical Report CRSC-TR08-08, NCSU, April, 2008.

[16] R. B. Couch and J. A. Kasel, *Immunity to influenza in man*, Ann. Rev. Microbiol., **31** (1983), 529–549.

[17] J. B. Cruz, Jr., ed., "System Sensitivity Analysis," Dowden, Hutchinson & Ross, Inc., Stroudsberg, PA, 1973.

[18] M. Davidian and D. M. Giltinan, "Nonlinear Models for Repeated Measurement Data," Chapman & Hall, Boca Raton, 1995.

[19] K. Dietz, *The estimation of the basic reproduction number for infectious diseases*, Stat. Methods Med. Res., **2** (1993), 23–41.

[20] J. Dushoff, J. B. Plotkin, S. A. Levin and D. J. D. Earn, *Dynamical resonance can account for seasonality of influenza epidemics*, Proc. Natl. Acad. Sci. USA, **101** (2004), 16915–16916.

[21] M. Eslami, "Theory of Sensitivity in Dynamic Systems: an Introduction," Springer-Verlag, New York, NY, 1994.

[22] N. M. Ferguson, A. P. Galvani and R. M. Bush, *Ecological and immunological determinants of influenza evolution*, Nature, **422** (2003), 428–433.

[23] A. Flahault, S. Letrait, P. Blin, S. Hazout, J. Menares and A. J. Valleron, *Modeling the 1985 influenza epidemic in France*, Stat. Med., **7** (1988), 1147–1155.

[24] L. Forsberg White and M. Pagano, *A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic*, Stat. Med., **27** (2008), 2999–3016.

[25] P. M. Frank, "Introduction to System Sensitivity Theory," Academic Press, New York, NY, 1978.

[26] H. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., **42** (2000), 599–653.

[27] M. Kleiber, H. Antunez, T. D. Hien and P. Kowalczyk, "Parameter Sensitivity in Nonlinear Mechanics: Theory and Finite Element Computations," John Wiley & Sons, Chichester, 1997.

[28] J. C. Lagarias, J. A. Reeds, M. H. Wright and P. E. Wright, *Convergence properties of the Nelder-Mead simplex method in low dimensions*, SIAM J. Optimiz., **9** (1999), 112–147.

[29] I. M. Longini, J. S. Koopman, A. S. Monto and J. P. Fox, *Estimating household and community transmission parameters for influenza*, Am. J. Epidemiol., **115** (1982), 736–751.

[30] A. L. Lloyd, *The dependence of viral parameter estimates on the assumed viral life cycle: Limitations of studies of viral load data*, Proc. R. Soc. Lond. B, **268** (2001), 847–854.

[31] H. Nishura, *Time variations in the transmissibility of pandemic influenza in Prussia, Germany, from 1918-19*, Theor. Biol. Med. Model., **4** (2007); Published online (http://www.tbiomed.com/content/4/1/20) DOI: 10.1186/1742-4682-4-20.

[32] M. Nuno, G. Chowell, X. Wang and C. Castillo-Chávez, *On the role of cross-immunity and vaccination in the survival of less-fit flu strains*, Theor. Pop. Biol., **71** (2007), 20–29.

[33] A. Saltelli, K. Chan and E. M. Scott, eds., "Sensitivity Analysis," John Wiley & Sons, Chichester, 2000.

[34] G. A. F. Seber and C. J. Wild, "Nonlinear Regression," John Wiley & Sons, Chichester, 2003.

[35] C. C. Spicer and C. J. Lawrence, *Epidemic influenza in greater London*, J. Hyg. Camb., **93** (1984), 105–112.

[36] W. Thompson, D. Shay, E. Weintraub, L. Brammer, N. Cox, L. Anderson and K. Fukuda, *Mortality associated with influenza and respiratory syncytial virus in the United States*, JAMA, **289** (2003), 179–186.

[37] US Department of Health and Human Services, website: http://www.pandemicflu.gov/general/historicaloverview.html, accessed on December 16, 2006.

[38] C. Viboud, T. Tam, D. Fleming, A. Handel, M. Miller and L. Simonsen, *Transmissibility and mortality impact of epidemic and pandemic influenza, with emphasis on the unusually deadly 1951 epidemic*, Vaccine, **24** (2006), 6701–6707.

[39] J. Wallinga and P. Teunis, *Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures*, Am. J. Epidemiol., **160** (2004), 509–516.

[40] H. J. Wearing, P. Rohani and M. Keeling, *Appropriate models for the management of infectious diseases*, PLoS Medicine, **2** (2005), e174.

Received April 24, 2008. Accepted August 4, 2008.

E-mail address: acintro@ncsu.edu
E-mail address: ccchavez@asu.edu
E-mail address: lmbett@lanl.gov
E-mail address: alun_lloyd@ncsu.edu
E-mail address: htbanks@ncsu.edu