Published in final edited form as:

J Theor Biol. 2017 May 07; 420: 68–81. doi:10.1016/j.jtbi.2017.01.032.

Model distinguishability and inference robustness in mechanisms of cholera transmission and loss of immunity

Elizabeth C. Lee¹, Michael R. Kelly Jr.², Brad M. Ochocki³, Segun M. Akinwumi⁴, Karen E. S. Hamre⁵, Joseph H. Tien², and Marisa C. Eisenberg⁶

¹Department of Biology, Georgetown University, 37th and O Streets, NW, Washington, DC 20057, United States of America

²Division of Natural Sciences and Mathematics, Transylvania University, 300 North Broadway, Lexington, KY 40508, and Department of Mathematics, The Ohio State University, 231 West 18th Ave, Columbus, OH 43210, United States of America

³Department of BioSciences, Program in Ecology and Evolutionary Biology, Rice University, 6100 Main Street, MS-170, Houston, TX 77005-1892, United States of America

⁴Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, T6G 2G1, Canada

⁵Division of Global Pediatrics and Division of Epidemiology and Community Health, University of Minnesota, 717 Delaware Street SE, 3rd Floor, Minneapolis, MN 55414, United States of America

⁶Departments of Epidemiology and Mathematics, University of Michigan, Ann Arbor, 1415 Washington Heights, Ann Arbor, 48109, USA

Abstract

Mathematical models of cholera and waterborne disease vary widely in their structures, in terms of transmission pathways, loss of immunity, and a range of other features. These differences can affect model dynamics, with different models potentially yielding different predictions and parameter estimates from the same data. Given the increasing use of mathematical models to inform public health decision-making, it is important to assess model distinguishability (whether models can be distinguished based on fit to data) and inference robustness (whether inferences from the model are robust to realistic variations in model structure).

In this paper, we examined the effects of uncertainty in model structure in the context of epidemic cholera, testing a range of models with differences in transmission and loss of immunity structure, based on known features of cholera epidemiology. We fit these models to simulated epidemic and long-term data, as well as data from the 2006 Angola epidemic. We evaluated model distinguishability based on fit to data, and whether the parameter values, model behavior, and forecasting ability can accurately be inferred from incidence data.

Correspondence to: Marisa C. Eisenberg.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

In general, all models were able to successfully fit to all data sets, both real and simulated, regardless of whether the model generating the simulated data matched the fitted model. However, in the long-term data, the best model fits were achieved when the loss of immunity structures matched those of the model that simulated the data. Two parameters, one representing person-to-person transmission and the other representing the reporting rate, were accurately estimated across all models, while the remaining parameters showed broad variation across the different models and data sets. The basic reproduction number (R_0) was often poorly estimated even using the correct model, due to practical unidentifiability issues in the waterborne transmission pathway which were consistent across all models. Forecasting efforts using noisy data were not successful early in the outbreaks, but once the epidemic peak had been achieved, most models were able to capture the downward incidence trajectory with similar accuracy. Forecasting from noise-free data was generally successful for all outbreak stages using any model.

Our results suggest that we are unlikely to be able to infer mechanistic details from epidemic case data alone, underscoring the need for broader data collection, such as immunity/serology status, pathogen dose response curves, and environmental pathogen data. Nonetheless, with sufficient data, conclusions from forecasting and some parameter estimates were robust to variations in the model structure, and comparative modeling can help to determine how realistic variations in model structure may affect the conclusions drawn from models and data.

Keywords

model misspecification; parameter estimation; model structure; comparative modeling

1 Introduction

Cholera is a waterborne disease caused by the bacterium *Vibrio cholerae*, which manifests as severe diarrhea and vomiting leading to dehydration. Left untreated, cholera can be up to 50% fatal, but rehydration treatment can greatly reduce case fatality rates to as low as 1% [1,2]. Worldwide, cholera causes three to five million cases and over 100,000 deaths per year [3]. Numerous mathematical models of cholera transmission have been proposed to investigate factors that impact the dynamics and transmission of waterborne diseases [4–13], and the ongoing cholera epidemic in Haiti has spurred additional interest in the subject [12,14–18].

Due to the range of indirect transmission pathways and timescales, which may be represented by environmental water sources, household water containers, foodborne transmission, and more [19,20], commonly used mathematical models for cholera vary widely in their level of detail, spatial scale, and model structure. Some models use a single term to represent a composite set of transmission mechanisms, while others include multiple timescales of transmission [4,8]. In addition to standard mass-action transmission [8,14], many models use nonlinear transmission functions for waterborne transmission [4,21] to reflect the dose response for cholera in the water. Models may also include an asymptomatic transmission pathway, a hyperinfectious state for the bacteria immediately after shedding, and other ecological and environmental factors in the environmental reservoir, such as effects of vibriophages, plankton, weather, and climate [4–6,8,10,18,22–26].

In addition to the variation in transmission mechanisms, loss of immunity to cholera is poorly understood and therefore, modeled with many different assumptions. Estimates of the length of immunity in the literature range widely from several months to three to ten years [6, 27, 28]. Immunity to cholera is of particular interest given the recent and ongoing oral cholera vaccine campaigns worldwide, including in Haiti, Bangladesh, and Thailand [29–32], which raise additional questions of how vaccine-derived immunity compares to immunity derived from infection.

As modeling gains prevalence among policy makers in public health [16,33–37], comparative or ensemble modeling approaches have been increasingly viewed as a way to ensure that the results of parameter estimation, forecasting efforts, and the evaluation of intervention strategies are conserved across the range of realistic model structures [38,39]. Two related concepts are useful to consider in these efforts—*model distinguishability* addresses whether candidate models can be distinguished by their fits to empirical data [40], and *inference robustness assessment* examine whether conclusions drawn from a particular model are robust to realistic variations in the assumptions and model structure [38]. Both of these concepts are important to evaluate in the model-building process when model results are used as the basis for decision making. While parameter uncertainty has been highlighted for cholera outbreaks both in general [26,41] and in Haiti-specific contexts [25,41], model distinguishability is a higher order examination of uncertainty at the model structure level. In previous work [20], we examined identifiability of a very simple but commonly used environmental transmission model of cholera, examining structural (theoretical) and practical identifiability in the context of distinguishing between direct (non-environmentally driven) and indirect (environmentally-driven) cholera transmission. However, the model considered there did not include a range of real world features known in cholera epidemiology (e.g. loss of immunity, dose response, asymptomatic cases) [41], and relatedly it did not consider issues of model distinguishability (beyond considering direct and indirect transmission), which are highly relevant given the real-world complexity and range of models in the literature.

In this paper, we examine the effects of uncertainty in model structure on cholera disease dynamics and inference by considering five models with different transmission and loss of immunity mechanisms. The models under consideration share a common base, the SIWR model of Tien and Earn [8]. The SIWR model is an extension of the classic Susceptible-Infected-Recovered (SIR) [42] with an added compartment for pathogen concentration in an aquatic reservoir (W) [8]. In addition to the person-to-person and water transmission pathways of the base SIWR model, we evaluate two additional transmission-related features: a Michaelis-Menten dose response term for waterborne transmission and an asymptomatic infection and transmission pathway [4,6,24,43]. We also consider three loss of immunity features not included in the SIWR model: exponential loss of immunity, stage-progression gamma-distributed loss of immunity, and a novel model that features progressively increasing susceptibility after recovery from infection.

Using these five deterministic SIWR-based models, we first simulate data from each model with different types of added noise. In the frame of model distinguishability and inference robustness assessment, we estimate model parameters and fit model data for all five models

to each simulated dataset and an empirical dataset from the 2006 cholera epidemic in Angola. We determine how well each model recaptures the underlying parameter and $\mathcal{R}_0$ values, as well as how each model fits to simulated data generated from different models. Finally, we forecast trajectories using parameters estimated from truncated simulated incidence data and compare model forecasts with that of the true simulated data.

2 Methods

2.1 Model descriptions

Here we introduce the five models of epidemic cholera disease dynamics that were used in our analyses. We base these models on the SIWR model of Tien and Earn [8], which includes two timescales of transmission: a fast, direct route (represented by β_I) which incorporates direct person-person transmission, foodborne and household transmission, and other non-environmental transmission pathways; and a slow, indirect route (β_W) which represents long-term transmission mediated by bacteria in the water. The scaled non-dimensional model equations are given by [8]:

$$\begin{aligned} s' &= \mu - \beta_I si - \beta_W sw - \mu s \\ i' &= \beta_W sw + \beta_I si - \gamma i - \mu i \\ w' &= \xi(i - w) \\ r' &= \gamma i - \mu r \end{aligned} \quad (1)$$

where s , i , and r represent the fractions of the population that are susceptible, infectious, and recovered, respectively. The w variable is proportional to the concentration of *V. cholerae* in the environment, and ξ represents the decay rate of the pathogen in the water (noting that the nondimensionalization in [8] rescales the water so that the parameter for pathogen shedding into the water is canceled and replaced with ξ , yielding the form of w seen above—see Supplementary Information for nondimensionalization details). Parameter definitions and values are given in Table 1. The basic reproduction number ($\mathcal{R}_0$) for the baseline model includes both transmission pathways and is given by

$$\mathcal{R}_0 = \frac{\beta_I + \beta_W}{\gamma + \mu}, \quad (2)$$

where γ is the recovery rate and μ is the natural death rate in the population. The five models in our study build upon this baseline SIWR model (Figure 1), with the aim of capturing a range of features commonly considered in loss of immunity and cholera dynamics. These features include: loss of immunity following exponential or gamma distributions, nonlinear dose response effects on transmission, and asymptomatic cases.

To connect these models with data, we use the same measurement (output) equation of $y = \kappa i$, as has been used with a range of data sets using the SIWR model [14, 18, 20] (including the same epidemic dataset as considered here). The parameter κ represents a combination of factors including the reporting rate, at-risk population size, potentially the infectious period

(if used as an approximation for incidence data), and potentially the symptomatic fraction (depending on the model considered), among others. We define $k = 1/\kappa$ and fit k rather than κ as this constrains the parameter estimation to be within $(0, 1)$ rather than $(1, \infty)$, making k easier to estimate [20].

Exponential model—The Exponential model builds on the baseline SIWR model by enabling recovered individuals to return to the susceptible state. We assume those in the recovered class (r) become susceptible (s) again after losing their immunity with rate parameter α . The duration of immunity among recovered individuals follows an exponential distribution. $\mathcal{R}_0$ for this model has the same structure as that of the baseline model (Eq. 2). The equations are given by:

$$\begin{aligned} s' &= \mu - \beta_I si - \beta_W sw - \mu s + \alpha r \\ i' &= \beta_W sw + \beta_I si - \gamma i - \mu i \\ w' &= \xi(i - w) \\ r' &= \gamma i - \mu r - \alpha r. \end{aligned} \quad (3)$$

Dose Response model—The Dose Response model adds to the Exponential model by accounting for nonlinear dose response effects in the waterborne transmission pathway [4, 41, 43, 44]. Challenge studies in which volunteers ingested cholera and other pathogens have shown that the probability of transmission is a nonlinear function of dose size, with sigmoidal or Hill function-shaped curves that show low probabilities of infection at low concentrations and saturating probability of infection once the bacteria attain sufficiently high concentrations in the aquatic reservoir. Following previous studies, this model incorporated a Michaelis-Menten dose response curve in the parameter for waterborne transmission [4, 43]. We assume susceptible individuals become infected at a rate $\frac{\beta_{Wmax} w}{K + w}$, where β_{Wmax} is the maximum force of infection from the water and K is the concentration of *V. cholerae* that gives a 50% chance of water-transmitted cholera infection. As in the Exponential model, individuals lose immunity at a rate α . The model equations are given by:

$$\begin{aligned} s' &= \mu - \beta_I si - \frac{\beta_{Wmax} w}{K + w} s - \mu s + \alpha r \\ i' &= \frac{\beta_{Wmax} w}{K + w} s + \beta_I si - \gamma i - \mu i \\ w' &= \xi(i - w) \\ r' &= \gamma i - \mu r - \alpha r \end{aligned} \quad (4)$$

and $\mathcal{R}_0$ for this model is given by:

$$R_0 = \frac{\beta_I + \frac{\beta_{Wmax}}{K}}{\gamma + \mu}.$$

Asymptomatic model—This model expands upon the Exponential model to include an asymptomatic infection pathway. Estimates of the ratio of asymptomatic to symptomatic cholera infections range from 3 to 100 [6]. We assumed 20% of infections were

symptomatic ($q = 0.2$) [41]. We assumed that all pathogens decay at the same rate ξ , regardless of which class they are shed from. Shedding rates differ by symptomatic/asymptomatic status [41], with differential shedding rates σ_S and σ_A . Scaling by these shedding rates and the other model parameters in the nondimensionalization (given in the Supplementary Information) results in the following model equations:

$$\begin{aligned}
 s' &= \mu - \beta_{IA} si_A - \beta_{IS} si_S - \beta_W sw - \mu s + \alpha_A r_A + \alpha_S r_S \\
 i_S' &= q(\beta_W sw + \beta_{IS} si_S + \beta_{IA} si_A) - \gamma i_S - \mu i_S \\
 i_A' &= (1-q)(\beta_W sw + \beta_{IS} si_S + \beta_{IA} si_A) - \gamma i_A - \mu i_A \\
 w' &= \xi(i_A + i_S - w) \\
 r_S' &= \gamma i_S - \mu r_S - \alpha_S r_S \\
 r_A' &= \gamma i_A - \mu r_A - \alpha_A r_A
 \end{aligned} \tag{5}$$

where $\mathcal{R}_0$ for this model is given by:

$$R_0 = \frac{1}{\gamma + \mu} [q(\beta_{IS} + \beta_W) + (1-q)(\beta_{IA} + \beta_W)].$$

Gamma model—As opposed to exponential loss of immunity, the Gamma model represents waning immunity through a chain of n recovered classes, which results in an immunity duration that has a gamma distribution (See [45] and [46] for examples). The number of classes depends on the assumed distribution on length of immunity; a greater number of compartments yields a duration that begins to approximate a time delay. In this model, individuals in the chain of recovered classes retain complete immunity until they return to the susceptible class, with equations given by:

$$\begin{aligned}
 s' &= \mu - \beta_I si - \beta_W sw - \mu s + n\alpha r_n \\
 i' &= \beta_W sw + \beta_I si - \gamma i - \mu i \\
 w' &= \xi(i - w) \\
 r_1' &= \gamma i - \mu r_1 - n\alpha r_1 \\
 &\vdots \\
 r_n' &= n\alpha r_{n-1} - \mu r_n - n\alpha r_n
 \end{aligned} \tag{6}$$

and $\mathcal{R}_0$ for this model is same as Equation 2. We note that the progression through each stage of immunity occurs at rate $n\alpha$ (where n is the number of compartments), so that the expected duration of immunity (α^{-1}) matched the baseline model.

Progressive Susceptibility model—Given the wide range of estimates for the length of immunity in cholera, we also developed a new model that allows for progressively increasing susceptibility after infection over time. This may more realistically reflect the dynamics of loss of immunity, which is likely to be progressive rather than switch-like. Similar models of progressive waning have been used in a range of other contexts [47], and a distribution of cholera antibody titers are often observed in cholera vaccine studies [48]. The Progressive Susceptibility model again represents loss of immunity through a chain of n

classes, but individuals are at least partially susceptible at all points in the chain. The transmission rates are given as functions of location in the chain (Figure 1), where more recently recovered individuals are less likely to become infected. As individuals progress through the chain, immunity to water-borne and person-person infection wanes, with susceptibility given by functions f and g , respectively. The forms of f and g were chosen so that the probability of infection is close to zero for individuals in the first susceptible class (s_n) and equal to one for individuals in the last susceptible class (s_0). The model equations are:

$$\begin{aligned}
 i' &= \sum_{j=1}^n \beta_I s_j i \cdot g(j) + \sum_{j=1}^n \beta_W s_j w \cdot f(j) - \gamma i - \mu i \\
 w' &= \xi(i - w) \\
 s_1' &= \mu + n\alpha s_2 - \beta_I s_1 i - \beta_W s_1 w - \mu s_1 \\
 &\vdots \\
 s_j' &= n\alpha s_{j+1} - n\alpha s_j - \beta_I g(j) s_j i - \beta_W f(j) s_j w - \mu s_j \\
 &\vdots \\
 s_n' &= \gamma i - n\alpha s_n - \beta_I g(n) s_n i - \beta_W f(n) s_n w - \mu s_n, \quad (7)
 \end{aligned}$$

where we take $f(j) = e^{-u_1(j-1)/(n-1)}$ and $g(j) = e^{-u_2(j-1)/(n-1)}$. For simplicity, we assumed baseline values of $u_1 = u_2 = n$, i.e. equal rates of waning immunity for waterborne and direct transmission. This form was chosen to make f and g equal to 1 for s_n and near zero for s_1 . We made the same adjustment to the waning immunity rates in each compartment as in the Gamma model, to yield the same expected duration of immunity. Again, $\mathcal{R}_0$ for this model is given by Equation 2.

2.2 Data simulation and baseline parameters

Model parameters were taken from previous modeling and empirical studies of epidemic cholera, primarily from the estimates in [20], which used a SIWR model fitted to the 2006 outbreak of cholera in Angola (Table 1). However, some updates to the parameter values were made, based on literature estimates and known biology, as described in more detail in the Supplementary Information (Section 6.3). For the epidemic scenarios, we simulated six sets of data from each model to represent several possible combinations of measurement error assumptions—noise-free, normal noise (noise-free model output as the mean; 10% of model output as the standard deviation), and Poisson noise (noise-free model output as the rate), and simulation durations—epidemic (100 days) and long-term (3 years). Our primary analyses were conducted with daily simulated data, and additional sensitivity analyses using weekly simulated data instead may be found in the Supplementary Information (Section 6.7). Throughout the paper, we use the term *simulation model* to refer to the model that simulated the data for the analysis. These 30 simulated datasets (five models $\times$ three noise options $\times$ two simulation durations) were used for both parameter estimation/model fitting and forecasting.

We used the average life expectancy of Angola, 55 years [49], to determine μ . Length of immunity to cholera (α) acquired after infection is not well understood, with estimates

ranging from weeks to months to years [6,27,28]. For our simulated data, we assumed immunity lasted one year. The complete set of values, descriptions, and sources for the baseline parameters used for data simulation is given in Table 1.

While most parameters were common across models, some models required additional, specialized parameters. For the Asymptomatic model, we chose β_{IA} and β_{IS} such that β_{IS} was four times greater than β_{IA} and the weighted average of β_{IA} and β_{IS} (where symptomatic infections represent $q = 20\%$ of the infected population) was equal to β_I for all of the other models. To compare β_I across models, in the Results for the Asymptomatic model we report this weighted average of the two transmission parameters. Reasoning that individuals with symptomatic infections will have longer lasting immunity than those with asymptomatic infections, we assumed that immunity from symptomatic infections (α_S) would last two years and then the faster loss of immunity from asymptomatic infections (α_A) was calculated to give an average rate equal to the other models. Similarly, the reported estimates of α for this model are given as the weighted average of α_S and α_A . Additionally, the k value used for the Asymptomatic model was adjusted based on the fact that only symptomatic cases are observed (Table 1), although the unadjusted form was reported to facilitate model comparison in the Results.

The Dose Response model parameter structure also deviated from the other models, with the waterborne transmission term expanded to form a Michaelis-Menten term. The half-saturation constant, K , was determined based on previous literature estimates [50]. To determine the value of the maximum transmission level, β_{Wmax} , there are in general several natural ways that one might match up the Michaelis-Menten function with the single waterborne transmission parameter β_W value used in the original SIWR model. For example, one might match this value to the linear approximation of the Michaelis-Menten function (i.e. let $\beta_W = \beta_{Wmax}/K$, which approximates the Michaelis-Menten function for low values of w), or to the maximum level of waterborne transmission (i.e. let $\beta_W = \beta_{Wmax}$), or to some intermediate value of transmission. The Dose Response model behaved quite similarly to the Exponential model when the linear approximation was used; we chose the saturated, maximum value of transmission to match the models, letting $\beta_{Wmax} = \beta_W$ for our baseline parameters, in order to accentuate the differences between models.

For the simulated data, we assumed simulation model initial conditions with 1% of the population infected and the remaining population susceptible. For the Asymptomatic model, we assumed that 0.2% of the population was I_S , 0.8% of the population was I_A , and the remaining population was susceptible. All model simulations and parameter estimation were done in MATLAB. We also evaluated the baseline $\mathcal{R}_0$ for each model using the baseline parameter values—with these parameters, all models have an $\mathcal{R}_0$ of 3, except for the Dose Response model, which has an $\mathcal{R}_0$ of 6.

2.3 Identifiability of parameters

Identifiability and uncertainty quantification methods allow us to address whether, and with what degree of certainty, it is possible to uniquely recover the parameters of a model for a given dataset. In general, two broad classes of identifiability are often considered—structural and practical identifiability [53]. Structural identifiability addresses whether it is possible to

estimate the parameters based purely on the structure of the model and the measurements, typically assuming a best-case of perfect, noise-free data. Practical identifiability considers how the amount and quality of the data (e.g. level of measurement error, number of replicates) may affect estimation of the parameters. One commonly used method of evaluating identifiability is via the Fisher Information Matrix (FIM), a symmetric matrix that represents the amount of information about the parameters that is contained in the data [54,55]. The rank of the FIM corresponds to the number of (locally structurally) identifiable parameters, and taking the inverse of the FIM gives an approximation of the covariance matrix for the parameters using the Cramer-Rao bound [53,54,56,57]. We calculated the FIM for each model (assuming the same parameters as for the model's simulated dataset, with daily data for one year), and evaluated both the rank and covariance matrix. From the parameter variances, we calculated the percent coefficient of variation (% CV) to evaluate parameter uncertainty. The % CV takes into account the size of the parameter value when evaluating uncertainty, with $\%CV = \frac{\sigma_p}{p} \times 100$ (where σ_p is the standard deviation and p is the value of the parameter). We used a tolerance value [58], where % CV = 100 indicates parameter identifiability. Typically, structurally unidentifiable parameters will have extremely large (e.g., numerically near-infinite, typically $> 10^6$ [58]) % CVs, while practically unidentifiable parameters have finite but large % CVs.

2.4 Parameter estimation, model fitting, and model distinguishability

To investigate model distinguishability, we considered whether epidemic fits were best when the data simulation model and fitting model matched versus when it did not. Throughout the paper, we use the term *fitting model* to refer to the model that was used to fit an epidemic dataset. To perform model fitting and parameter estimation from the simulated data, we used two sets of parameters to initialize the optimization: *informed starting parameters*, which were the true parameters from the simulation model (i.e. the parameters used in generating the simulated data, given in Table 1), and *naïve starting parameters*, which were randomly chosen within $\pm 20\%$ of the true parameters. Using maximum likelihood estimation and Nelder-Mead optimization in MATLAB (fminsearch function), we fit each of the five model structures to the 30 simulated datasets. When fitting our models to data, we estimated all parameters except the birth/death rate, μ , and the recovery rate, γ , which are known [14,49,52]. To compare model fits to simulated data, we calculated and compared Akaike information criterion (AIC) values for the models [59]; for ease of comparison, we report

AIC values, defined as the difference in AIC between the model of interest and the best fit model (i.e., model with the lowest AIC) for a given dataset. To compare parameter estimates among the fitting models, we examined the percent deviation from the true value, which was calculated as the difference between the estimate and the true value, divided by the true value. In this descriptive analysis, a deviation less than 20% of the true value was considered 'accurately recaptured' because the variation should be smaller than the potential deviation in starting parameters. Additional details on the parameter estimation methods and optimizer settings are also given in the Supplementary Information.

For model initial conditions when fitting, we let $\tilde{x}(0) = kx(0)$, where $x(0)$ is the initial data value, and then assumed the remaining population to be susceptible. For the Asymptomatic model, we assumed that the observed infected were symptomatic, and generated an

additional fraction of asymptotically infected individuals based on the proportionality parameter q (with the remainder of the population again assumed susceptible).

2.5 Forecasts of epidemic data

In addition to parameter estimation, we evaluated the ability for the study models to forecast epidemic trajectories when provided with truncated and simulated epidemic data. Using 10, 30, and 50 days of simulated incidence data (out of 100-day simulated epidemic data, capturing the early phase, before the peak, and during/after the peak) with no noise and normal noise from the five study models, we estimated parameters from each fitting model. These parameter estimates were then used to generate a trajectory forecast for the remainder of the 100-day epidemic. Throughout the study, we use *forecasting model* to refer to the model used to estimate these parameters and generate the forward-looking trajectory. For the parameter estimates in this section, we used the naive starting values for the no noise forecasts and both naive and informed starting values for the normal noise forecasts.

2.6 Application to 2006 cholera outbreak in Angola

To evaluate these distinguishability questions in practice, we investigated the ability for the five study models to fit and forecast weekly data from an empirical cholera epidemic. We applied the same methods of parameter estimation, model fitting, and forecasting to data from the recent 2006 cholera outbreak in Angola (data courtesy of the WHO Cholera Task Force), shown in Figure 6. The outbreak began in Luanda province (February 13, 2006), and eventually spread to 16 of 18 provinces, resulting in 82,204 cases and 3,092 deaths [20,60]. The ‘true’ parameter values are unknown, so we considered both the ‘informed’ and ‘naive’ sets of starting parameter values used in the simulated data for fitting to real data as well. In this case, the ‘informed’ starting parameter values of course do not represent the true parameter values, but are in some sense more ‘informed’, as the baseline values given in Table 1 are based on the best fit parameters for the the original SIWR model to this data set in [20]. For fitting, we used weighted least squares assuming variance proportional to the data, as previously used in [20], with further details Supplementary Information Section 6.3. The same initial condition setup was used in these fits as for fitting to the simulated data.

3 Results

3.1 Identifiability of parameters

The FIMs for all models except for the Asymptomatic model were full rank. All FIM’s were numerically invertible, with the percent coefficients of variation (% CVs) shown in Table 2. Apart from the aforementioned rank deficiency in the Asymptomatic model, all parameter % CVs were small enough that the parameters would be considered structurally identifiable. The Asymptomatic model transmission parameters can be shown analytically and numerically to be structurally unidentifiable (Supplementary Information). Thus, for the Asymptomatic model, we report the identifiable combination $\beta_I = q\beta_{IS} + (1 - q)\beta_{IA}$ in our subsequent analyses, rather than the individual unidentifiable parameters β_{IS} and β_{IA} (Supplementary Figure S1).

3.2 Parameter estimation and model fitting to epidemic data

For the 100-day simulation duration, across the three noise cases, all five models fit the simulated data well (Figure 2 and Supplementary Figure S5). Nearly all model fits matched the simulated noisy data closely throughout the trajectory, with few runs (strings of consecutive data points on one side of the fit) of obviously correlated residuals. Correspondingly, the AIC values were nearly indistinguishable across model fits, but the Exponential or Dose Response model fits consistently had the lowest (i.e., best) numerical values in 25 out of 30 cases, while the Asymptomatic and Progressive Susceptibility model fits frequently matched or followed close behind (Supplementary Tables S1 – S6).

The true parameter values of direct transmission (β_I) and measurement scaling (k) were recovered much more accurately and precisely than the other parameters that were common across all models (Figure 3), and neither of these parameters had identifiability issues in our study (Table 2) or a previous cholera modeling study [20]. For β_I , the magnitude in parameter estimate deviations across all five fitting models was under 20%. For k , the magnitude in parameter estimate deviations only exceeded 20% in a few instances when fitting to normal noise data.

In congruence with a previous study [20], we found that β_W and ξ formed a practically unidentifiable combination in all models (Figure 3 and Supplementary Figure S2), which might explain their large deviations from the true parameter value and the apparent tradeoffs between the two parameters. We also considered that some of the deviations in β_W (and by extension ξ), may be due to the different choice of structure for β_W in the Dose Response model. We evaluated the parameter variation without the Dose Response parameter estimates and simulated data, and found that while there was some attenuation of the more extreme outliers for the estimates of β_W across all models, the remaining parameter distributions remained largely identical (shown in Supplementary Figure S2).

As expected, parameter estimate recapture accuracy improved as simulated data decreased in noise (Figure S3). Parameter estimates for all combinations of simulation model data and fitting models are reported in Supplementary Tables S1 – S6. Additionally, the parameter estimates and fits using weekly rather than daily reporting were quite similar to the daily data in terms of parameter estimation accuracy (although variance in deviations was slightly larger for some parameters) and yielded essentially the same fits (visually) as the daily data fits (Supplementary Information Section 6.7).

3.3 Parameter estimation and model fitting to long-term data

For the 3-year data, all models were able to fit the qualitative dynamics of the simulated data, and with a few exceptions, visually all models matched all data sets well (Supplementary Figures S6 – S7). However, unlike the 100-day data, the lowest AIC values were most often achieved when the same model which generated the data was used for fitting (though often other models yielded very close AIC values, with differences ≤ 5). AIC values were consistently lower when models where the waning immunity distributions matched those of the data simulation model (exponentially-distributed waning immunity: Exponential, Dose Response, and Asymptomatic models; gamma-distributed waning

immunity: Gamma and Progressive Susceptibility models) (Supplementary Tables S7 – S12). Similarly, estimates from models where the waning immunity distributions matched those of the data simulation model were more accurate (Supplementary Figure S4). We note that our ability to distinguish model fits through large differences in AIC was facilitated by the large volume of simulated data—daily data over three years. In practice, time series data would be much sparser, making it more difficult to distinguish the different models. However, as with the 100-day data, moving to weekly rather than daily data reporting did not substantially alter the results (Supplementary Information Section 6.7).

The parameter estimate accuracy did not appear to differ strongly as a function of the type of added noise (Supplementary Figure S4). As with the 100-day data analyses, β_I was the most accurate estimated parameter across all models, followed by k , although k estimates appeared to be biased when the waning immunity distribution of the data simulation model did not match that of the fitting model (Figure 3, Supplementary Tables S7 – S12). In general, α was underestimated for gamma-distributed waning models fitted to exponentially-distributed waning models, and overestimated in the reverse case. The effect of the practically identifiable combination between β_W and ξ was also present in the 3-year data, with underestimates of β_W tending to be compensated by overestimates of ξ (Figure 3, Supplementary Tables S7 – S12).

3.4 Forecasts from simulated epidemic data

Epidemic trajectories were projected forward based on 10, 30, and 50 days of observed, simulated normal noise data (early in the epidemic, before the peak, and at the peak, respectively) for each study model. For the normal noise data, all forecasting models generated trajectories that were poorly representative of the true simulated data when only 10 or 30 data points were observed (Figure 4, Supplementary Figures S11). Even with informed starting parameters, even when forecasting models matched simulation models, these truncated epidemic datasets did not allow for accurate forecasts. While forecasts generated from 30 observations as compared to 10 observations were improved, performance remained poor overall, with only one out of 25 forecasts able to accurately capture both the timing and magnitude of the simulated epidemic peak (Exponential model forecasting Dose Response simulated data). In fact, 11 of 25 forecasts continued to project exponential growth upwards past the 100-day simulation period. Forecasts generated from 50 observed data points were significantly improved, with 15 of 25 forecasts roughly capturing the downward trajectory of the true data. As with the previous forecasts, a match between the simulation and forecasting model did not necessarily generate the most accurate forecast. Forecasts were also generated with naive starting parameters using observations with no added noise (Supplementary Figure S12), and in this case much higher forecasting accuracy was observed. Many models were able to approximately predict the model dynamics from only the 10-day data, and all models correctly forecasted the epidemic trajectory once 30 or 50 days of noise-free data were included.

The estimated parameters were also compared to the true parameters used to generate the forecasting data set (Figure 5, Figure S13). The direct transmission parameter β_I was the only one consistently recovered across varying levels of observation, although other

parameters showed some small improvements in accuracy as the amount of data increased. As might be expected, parameter estimation was much more accurate and precise when forecasting from data with no added noise as compared to data with added normal noise.

3.5 Application to 2006 cholera outbreak in Angola

We now turn our attention to data collected from a 2006 cholera outbreak in Angola. The fits of the models to the data are shown in Figure 6. The corresponding parameter estimates and AIC values for these fits, accounting for both informed and naive starting parameters, are given in Table 3. All models fit the data well, with similar trajectories that captured the overall shape, peak, and duration of the epidemic. The lowest %CVs were generally (though not always) found in the parameters k and β_I for both the ‘informed’ and ‘naive’ starting parameters (where ‘informed’ here indicates the values chosen from the literature for data simulation, although these are not the true values as they were for the simulated data). We note though that FIM-based variances are asymptotic, lower bounds of the variance, based on the curvature of the likelihood locally at a single point in parameter space (the parameter estimates). Thus, they often do not capture issues of practical unidentifiability [61], such as the β_W - ξ tradeoff observed with the simulated data and in previous work with the Angola data [20]. The Progressive Susceptibility model gave the lowest AIC for both naive and informed starting parameters (3038 and 2876, compared to AICs greater than 3400 for all other models), suggesting that the Progressive Susceptibility model provides the best fit of the Angola data. Interestingly, this model showed very little direct transmission (β_I), suggesting that waterborne transmission may be enough to explain the outbreak.

4 Discussion

Modeling (in cholera and more broadly) is often used for inference/understanding mechanism, as well as for prediction of system dynamics, outcomes, and effects of interventions. Distinguishability and identifiability play key roles in these aims: inferences about mechanism may be incorrectly drawn from an unidentifiable model or from a misspecified model that is indistinguishable from the true mechanism. Predictions and intervention projections from such models can similarly go awry; on the other hand, equivalent models or parameter sets may yield the same predictions of measured variables. The complex interplay between parameter uncertainty and model uncertainty thus has direct effects on our ability to draw useful conclusions from models—for cholera in particular, there are a wide range of realistic mechanistic features that might be included, with significant potential for model and parameter uncertainty.

Here, we have examined how model uncertainty and misspecification affect parameter estimates, model fits, and model forecasting ability. We considered five deterministic SIWR-based model structures, each including different hypothesized mechanisms of cholera transmission and loss of immunity, using both simulated data and data from the 2006 cholera epidemic in Angola. We found that goodness-of-fit criteria were unable to distinguish misspecified model fits from those of the true model for simulated epidemic data, but model fits for long-term data were found to be specific to the type of waning immunity that was implemented (i.e., exponential- vs. gamma-distributed loss of immunity mechanisms).

Moreover, some parameters were consistently estimated across all models and datasets, suggesting that it may be possible to estimate some parameters even when the underlying mechanistic model is unknown, though further work to evaluate the generalizability of this result is needed. We found a consistent issue of practical unidentifiability between the transmission and pathogen decay parameters, which results in poor estimation of $\mathcal{R}_0$ across all models, and may persist as an issue for environmental transmission models more broadly. Forecasting from noisy data was generally poor unless the peak was captured, even when the model that generated the data and the model that produced the forecast were the same. However, if noise-free data was used, forecasting was accurate using all models even with only very short data, highlighting the importance of accurate surveillance when using models for forecasting.

We showed that each model was able to capture the epidemic trajectories for all datasets, real and simulated, regardless of the source of the simulated data, type of noise, or starting parameters. Indeed, often the best fit to a given epidemic dataset did not come from the model which generated it. This suggests that it may not be possible to infer the true underlying mechanisms of loss of immunity and transmission solely from epidemic cholera incidence or prevalence data. This highlights the distinction between the issues of identifiability (parameter uncertainty) and distinguishability (model uncertainty)—all model parameters were structurally identifiable (excepting the parameter combination for the Asymptomatic model), but the model structure itself may not be possible to infer from the data.

When a longer, 3-year data set was considered—as noted above, in this case it was generally possible to distinguish exponentially-distributed waning models (Exponential, Dose Response, and Asymptomatic) from gamma-distributed waning models (Gamma and Progressive Susceptibility). Indeed, the best-fit model to the simulated 3-year data was often the model which generated it, although typically the other models with the same waning distribution had very similar AIC values (differences ≤ 5), such that the models may not be distinguishable in practice. These results suggest that some degree of model selection, at least for loss of immunity mechanisms, may be possible with longer term data sets. Other additional data might add to the distinguishability of realistic models, such as data on pathogen shedding or decay rates, bacterial concentration in water sources, dose response curves, asymptomatic fraction, or immunity/serology data such as antibody titers. Such data could constrain the associated mechanistic parameters and improve identifiability of other parameters (e.g. measuring the decay rate ξ could resolve the unidentifiability issues with β_W), or make the distinguishability question somewhat moot if incorporating the additional data entails adding the mechanism in question (e.g. using dose response data might require including dose response as a mechanism in the model).

Our study design also enabled the systematic assessment of parameter estimates under conditions of model misspecification. As noted above, all parameters were structurally identifiable (apart from the identifiable combination for β_{IA} and β_{IS}), but unfortunately, identifiability analyses only inform the validity of parameter estimates when the true underlying model is known. When we consider the variation of the parameter estimates across different model misspecification scenarios, the picture is more complex. As β_I and k

were recovered quite well for all simulated data sets, this suggests that the “fast” transmission parameter and reporting rate/population factor may be accurately estimated from any of the common SIWR model variants, even if the true underlying disease mechanisms are unknown. Interestingly, the combined parameter for reporting rate and at-risk population size (k) was more accurately recovered from the misspecified models when the 100-day data sets were used rather than the 3-year data (Figure 3), suggesting that when model misspecification is present, more data may not always be better. It is likely that the bias in k helped to correct the fitting issues that resulted from incorrect model structure, but this would be difficult to assess in practice with real data.

Previous work noted that in the SIWR model, the parameters for waterborne transmission (β_W) and pathogen shedding in the water (ξ) became practically unidentifiable when noisy data was considered, and formed a practically identifiable combination [20]. It was conjectured there that this practical unidentifiability tradeoff between transmission and pathogen persistence may extend more broadly to many models that include a waterborne or environmental transmission pathway. Indeed, here we confirmed that β_W and ξ are practically unidentifiable and form a practically identifiable combination across multiple SIWR-based models. This underscores the need for improved data collection regarding waterborne cholera transmission in order to better inform modeling and public health intervention efforts. However, we note that while the two parameters were practically unidentifiable in all models, the curve of the identifiable combination traced in Figure 3 was quite similar across all models and data sets. This suggests that while these two parameters may individually be practically unidentifiable in all models, it may be possible to accurately estimate their combination across models and datasets in a manner similar to β_I and k . However, the practical unidentifiability of β_W also has important practical consequences for the usefulness of environmental transmission models in calculating $\mathcal{R}_0$. Due to this practical unidentifiability of β_W , estimates of $\mathcal{R}_0$ varied widely across all models, making them potentially unreliable estimates for the epidemic data in Angola and for all types of simulated data. Given that this unidentifiability has shown itself to be quite common across multiple models, this suggests that estimating $\mathcal{R}_0$ in practice for environmentally transmitted diseases may be challenging.

Estimates of α were fairly poor across both 100-day and 3-year scenarios, but the variances in the deviation from the true parameter were notably higher in the 100-day scenarios. The likely reason for this difference is that the 100-day period is too short to witness the effects of loss of immunity mechanisms. Consequently, parameter estimation and modeling for epidemic (as opposed to endemic) cholera dynamics may choose to exclude loss of immunity mechanisms in favor of a more parsimonious model structure since this parameter will not be well-estimated under these circumstances (e.g. as in [8,14] and others). In general, in the 3-year datasets, α was underestimated for gamma-distributed waning models fitted to exponential waning models, and overestimated in the reverse case. This may be because the gamma-distributed waning models used a slower loss of immunity to generate the flat trajectories after the initial peak, while the exponential-distributed waning models may have used a more rapid α to generate the subsequent peaks in the disease trajectory.

Given the possibility that daily reporting may give a more optimistic picture of the uncertainty on the parameters than would be possible with real data, we also examined weekly data (both simulated and using the Angola epidemic data). The results were quite similar to the daily data in terms of parameter estimation accuracy (although variance in deviations was slightly larger for some parameters) and yielded essentially the same fits (visually) as the daily data fits. The uncertainty in the estimates appears to derive from issues of practical identifiability that are either more fundamental to the models (e.g. the β_W - ξ tradeoff), or come from the data length rather than frequency (e.g. large uncertainty on α when the data length is less than one year).

In extending our models to empirical Angola epidemic data, we found that the new Progressive Susceptibility model presented here had the best absolute fit, capturing both the peak and tails of the Angola data. As previously noted, however, this does not provide confirmatory evidence that waning immunity mechanisms were the true driver of these epidemic dynamics, as other mechanisms might also produce this pattern, e.g. spatial dynamics as the epidemic spread across the country or reporting delays. These and other unknown, real-world processes that afflict empirical data may also explain the wider variation in model AICs as compared to the simulated data. Longitudinal data on sequential infections of cholera or on antibody titers over time are needed to validate the idea that recently infected individuals regain susceptibility progressively after recovery (and the rates at which they do so). Across all model fits to the Angola data, we found that waterborne transmission was consistently identified as the dominant mode of transmission, corroborating findings from a prior study using this data [20]. However, the identifiability issue with β_W (discussed above) may cast some doubt on the robustness of this finding, and may explain why estimates of $\mathcal{R}_0$ varied widely across all models. Future extensions of our model comparison framework to real epidemics should explore whether certain models or model features present systematic patterns of low or high estimates of $\mathcal{R}_0$.

Our results also underscore the importance of accurate surveillance data to forecasting efforts. For noisy data, only projections that took place after the epidemic peak appeared to capture any semblance of disease prevalence; projections were also poor when the forecasting model matched the simulation model, perhaps suggesting that it may not be possible to forecast accurately with deterministic SIWR models until the latter half of the epidemic. This is to be expected in the very early part of the epidemic, during the exponential growth phase, as during this phase the epidemic is largely linear on a log-scale and so can be explained by only two parameters. This idea is borne out in other real-time forecasting efforts—for example, efforts to forecast the trajectory of the 2014 Ebola epidemic in West Africa met with difficulty [62–65]. We note that our results illustrate how agreement across a range of models may not guarantee accuracy; however, once the epidemic peak was observed, all models tended to converge on similar and more accurate forecasts. Additionally, forecasting appeared to be quite promising at all outbreak stages among simulated data with no added noise. Parameter estimates from forecasting tended to improve with both decreasing noise and longer data sets. Together, these results suggest that if high quality or sufficient data are obtained, forecasting is possible even if the model structures do not match the underlying mechanisms; we suggest that accurate and timely surveillance during the early phase of an outbreak may substantially improve model

forecasting ability. Future work may examine whether early outbreak forecasting, discrimination of model structures, and inference from forecasting comparisons may be aided by additional data on weather and climate variables such as rainfall or sea surface temperature [18, 25, 66–68].

There are several limitations and future directions for this work, to consider different model structures, frameworks, and empirical datasets. Alternative cholera transmission mechanisms (such as hyperinfectivity [5,10,22,24,69]), spatial structure [14,17,70,71], and climate drivers [18,25,66–68] were also realistic features that could have been added to the suite of model structures we considered. Model comparison with hyperinfectivity and two potential environmental reservoirs were considered in study by Rinaldo et al. [25], using a larger, detailed spatial model of the 2010 Haiti cholera epidemic (including human and pathogen movement, river networks, climate drivers, etc.). While we only examined individual mechanisms added to the SIWR model here, more realistic combinations would be useful to consider. In particular, combining asymptomatic cases and waning immunity would be quite interesting as a next step. However, the lack of distinguishability between the individual mechanisms shown here suggests that models combining multiple features may encounter identifiability issues. We also did not evaluate our conclusions outside of the deterministic ODE framework, although partial differential equation (PDE), agent-based, and stochastic models are also frequently used in similar contexts [6, 68, 72, 73]. In addition, most of the parameter values were motivated from a single outbreak dataset (Angola, 2006), and may not reflect the breadth of parameter space for cholera epidemics. Additional work to further examine the identifiability, uncertainty of these models, and to test alternative parameter estimation methods (e.g. Bayesian and/or global approaches) would be warranted as well. For example, broader comparative analysis of the practical identifiability of different plausible models under different parameter regimes would be a natural next step. It should also be noted that only local optimization methods were used here, so further work examining global methods and explicitly examining local vs. global minima would be useful. The distinguishability analyses conducted here could potentially be repeated over a range of points in parameter space to examine how model distinguishability varies across this space. It would also be useful to consider alternate data assumptions (e.g. irregular data frequencies, delayed reporting, and additional noise and error assumptions); sensitivity analyses with weekly data frequency yielded parameter estimation and fit accuracy similar to that of the daily data (Supporting Information, Section 6.7). Further, here we only considered the effects of model structure on parameter estimation and forecasting. Additional work could use model comparison to evaluate the effects of model and parameter uncertainty on the optimal control of interventions, similar to that proposed by Akman and Schaefer [26]. Their focus on the optimal balance of vaccination and sanitation strategies may be particularly interesting, as the lack of distinguishability between models may be less problematic if multiple models yield consistent control strategies.

In conclusion, our study presents a systematic framework for the comparison of model mechanisms and inference that can be applied broadly in infectious disease epidemiology. Modeling requires simplifying assumptions to be an effective tool, but we show here that the choice of simplifications can have significant effects on the parameter estimates, and that it may be difficult or impossible to use fit to data as a criteria to distinguish between plausible

simplified models. Of course, in real-world contexts, there is unlikely to be a ‘true’ model in the sense that we considered in our simulated data—instead, each of the models represents a simplification, albeit with realistic elements. We note that our results also show that the most realistic model—a ‘super-model’ containing all of the elements considered individually across the five models here—would almost certainly be unidentifiable from case data alone. Public health decision makers desire tools that demonstrate the sensitivity of results to model assumptions (on the analysis side) and policy decisions (on the implementation side). Attempts to address the divide between decision maker needs and modelers’ ability to present useful, actionable information are on the rise, with the development of adaptive management, cost-benefit, and game theoretic frameworks in public health [74–80]. In demonstrating that goodness-of-fit is not sufficient to infer transmission and immunity mechanisms in a disease system, and that some parameter estimation may be valid even when mechanisms remain elusive, we similarly seek to inform modelers and decision makers about the relative utility of different analyses for specific types of policy questions. Our research raises questions about the appropriateness and equality of model structures and types for epidemiological inference, and future work should seek to further characterize the sensitivity of public health decision making in the context of model choice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This paper began as a group project at the joint NIMBioS-MBI-CAMBAM Graduate Summer School in 2013 (http://www.nimbios.org/education/WS_grad2013)—many thanks to Suzanne Lenhart, Paul Hurtado, Ariel Cintron-Arias, and the other co-organizers and presenters. The summer school was hosted at the National Institute for Mathematical and Biological Synthesis (NIMBioS), an Institute sponsored by the National Science Foundation through NSF Award DBI-1300426, with additional support from The University of Tennessee, Knoxville. NIMBioS hosted the summer school jointly with Mathematical Biosciences Institute (MBI) at The Ohio State University (National Science Foundation grant DMS 68 0931642) and the Centre for Applied Mathematics in Bioscience and Medicine (CAMBAM) at McGill University. The NIMBioS summer school provided support to ECL, MRK, BO, SMA, KM, and MCE, and SMA acknowledges partial travel support from PIMS IGTC, Canada for summer school attendance. We thank the World Health Organization Cholera Task Force for providing us with data from the 2006 cholera outbreak in Angola. This work was supported by the National Science Foundation grant OCE-1115881 (to JHT and MCE) and the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U01GM110712 (to MCE). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank three anonymous reviewers for their helpful comments on this manuscript.

References

1. Sack D, Sack R, Nair G, Siddique A. Cholera. *Lancet*. 2004; 363:223–233. [PubMed: 14738797]
2. WHO. World Health Organization - Cholera Factsheet. 2011. <http://www.who.int/mediacentre/factsheets/fs107/en/index.html>
3. World Health Organization (WHO). Cholera fact sheet n-107. 2010. <http://www.who.int/mediacentre/factsheets/fs107/en/index.html>
4. Codeco C. Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. *BMC Infectious Diseases*. 2001; 1
5. Hartley D, Morris J, Smith D. Hyperinfectivity: a critical element in the ability of *V. cholerae* to cause epidemics? *PLoS Medicine*. 2006; 3:63–69.
6. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. *Nature*. 2008; 454:877–880. [PubMed: 18704085]

7. Joh RI, Wang H, Weiss H, Weitz JS. Dynamics of Indirectly Transmitted Infectious Diseases with Immunological Threshold. *BULLETIN OF MATHEMATICAL BIOLOGY*. May.2009 71:845–862. [PubMed: 19096894]
8. Tien JH, Earn DJD. Multiple transmission pathways and disease dynamics in a waterborne pathogen model. *Bulletin of Mathematical Biology*. 2010; 72(6):1506–1533. [PubMed: 20143271]
9. Mwasa A, Tchuente JM. Mathematical analysis of a cholera model with public health interventions. *Biosystems*. 2011; 105(3):190–200. [PubMed: 21540075]
10. Shuai Z, Tien JH, van den Driessche P. Cholera models with hyperinfectivity and temporary immunity. *Bulletin of Mathematical Biology*. 2012; 74(10):2423–2445. [PubMed: 22864877]
11. Tian JP, Wang J. Global stability for cholera epidemic models. *Mathematical biosciences*. 2011; 232(1):31–41. [PubMed: 21513717]
12. Andrews JR, Basu S. Transmission dynamics and control of cholera in haiti: an epidemic model. *The Lancet*. 2011; 377(9773):1248–1255.
13. Sanches R, Ferreira C, Kraenkel R. The role of immunity and seasonality in cholera epidemics. *Bulletin of Mathematical Biology*. 2011; 73(12):2916–2931. [PubMed: 21468779]
14. Tuite AR, Tien J, Eisenberg M, Earn DJ, Ma J, Fisman DN. Cholera epidemic in Haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions. *Annals of Internal Medicine*. 2011; 154(9):593–601. [PubMed: 21383314]
15. Chao DL, Halloran ME, Longini IM Jr. Vaccination strategies for epidemic cholera in Haiti with implications for the developing world. *Proceedings of the National Academy of Sciences USA*. 2011; 108(17):7081–7085.
16. Abrams JY, Copeland JR, Tauxe RV, Date KA, Belay ED, Mody RK, Mintz ED. Real-time modelling used for outbreak management during a cholera epidemic, haiti, 2010–2011. *Epidemiology and Infection*. 2012;1–10.
17. Bertuzzo E, Mari L, Righetto L, Gatto M, Casagrandi R, Blokesch M, Rodriguez-Iturbe I, Rinaldo A. Prediction of the spatial evolution and effects of control measures for the unfolding haiti cholera outbreak. *Geophysical Research Letters*. 2011; 38:L06403.
18. Eisenberg MC, Kujbida G, Tuite AR, Fisman DN, Tien JH. Examining rainfall and cholera dynamics in haiti using statistical and dynamic modeling approaches. *Epidemics*. 2013; 5:197–207. [PubMed: 24267876]
19. Swardlow D, Mintz E, Rodriguez M, Tejada E, Ocampo C, Espejo L, Greene K, Saldana W, Seminario L, Tauxe R, Wells J, Bean N, Ries A, Pollack M, Vertiz B, Blake P. Waterborne transmission of epidemic cholera in Trujillo, Peru: lessons for a continent at risk. *Lancet*. 1992; 340(8810):28–32. [PubMed: 1351608]
20. Eisenberg M, Robertson S, Tien J. Identifiability and estimation of multiple transmission pathways in waterborne disease. *Journal of Theoretical Biology*. 2013; (324):84–102.
21. Mukandavire Z, Liao S, Wang J, Gaff H, Smith DL, Morris JG. Estimating the reproductive numbers for the 20082009 cholera outbreaks in zimbabwe. *Proceedings of the National Academy of Sciences*. 2011; 108(21):8767–8772.
22. Pascual M, Koelle K, Dobson AP. Hyperinfectivity in cholera: a new mechanism for an old epidemiological model? *PLoS Medicine*. 2006; 3(6):e278–281. [PubMed: 16789801]
23. Koelle K, Rodo X, Pascual M, Yunus M, Mostafa G. Refractory periods and climate forcing in cholera dynamics. *Nature*. 2005; 436:696–700. [PubMed: 16079845]
24. Miller-Neilan RL, Schaefer E, Gaff H, Fister KR, Lenhart S. Modeling optimal intervention strategies for cholera. *Bulletin of Mathematical Biology*. 2010; 72(8):2004–2018. [PubMed: 20204710]
25. Rinaldo A, Bertuzzo E, Mari L, Righetto L, Blokesch M, Gatto M, Casagrandi R, Murray M, Veseneckh SM, Rodriguez-Iturbe I. Reassessment of the 2010–2011 Haiti cholera outbreak and rainfall-driven multiseason projections. *Proceedings of the National Academy of Sciences USA*. 2012; 109(17):6602–6607.
26. Akman O, Schaefer E. An evolutionary computing approach for parameter estimation investigation of a model for cholera. *Journal of biological dynamics*. 2015; 9(1):147–158. [PubMed: 25948150]
27. Levine M, Black R, Clements M, Cisneros L, Nalin D, Young C. Duration of infection-derived immunity to cholera. *Journal of Infectious Diseases*. 1981; 143(6):818–820. [PubMed: 7252264]

28. Koelle K, Pascual M. Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. *The American Naturalist*. 2004; 163(6):901–913.
29. Tohme R, François J, Wannemuehler K, Iyengar P, Dismar A, Adrien P, Hyde T, Marston B, Mintz E, Katz M, et al. Oral cholera vaccine coverage, barriers to vaccination, and adverse events following vaccination, haiti, 2013 (1). *Emerging infectious diseases*. 2015; 21(6):984–991. [PubMed: 25988350]
30. O'Leary M, Mulholland K. Oral cholera vaccines in endemic countries. *The Lancet*. 2015; 386(10001):1321–1322.
31. Deen J, von Seidlein L, Luquero FJ, Troeger C, Reyburn R, Lopez AL, Debes A, Sack DA. The scenario approach for countries considering the addition of oral cholera vaccination in cholera preparedness and control plans. *The Lancet Infectious Diseases*. 2016; 16(1):125–129. [PubMed: 26494426]
32. Phares CR, Date K, Travers P, Déglise C, Wongjindanon N, Ortega L, Bhuket PRN. Mass vaccination with a two-dose oral cholera vaccine in a long-standing refugee camp, thailand. *Vaccine*. 2016; 34(1):128–133. [PubMed: 26549363]
33. Moyer VA. Screening for lung cancer: Us preventive services task force recommendation statement. *Annals of internal medicine*. 2014; 160(5):330–338. [PubMed: 24378917]
34. Auchincloss AH, Roux AVD. A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. *American journal of epidemiology*. 2008; 168(1):1–8. [PubMed: 18480064]
35. Grad YH, Miller JC, Lipsitch M. Cholera modeling: challenges to quantitative analysis and predicting the impact of interventions. *Epidemiology (Cambridge, Mass)*. 2012; 23(4):523.
36. Lofgren ET, Halloran ME, Rivers CM, Drake JM, Porco TC, Lewis B, Yang W, Vespignani A, Shaman J, Eisenberg JN, et al. Opinion: Mathematical models: A key tool for outbreak response. *Proceedings of the National Academy of Sciences*. 2014; 111(51):18095–18096.
37. Lipsitch M, Finelli L, Heffernan RT, Leung GM, Redd SC. for the 2009 H1N1 Surveillance Group. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza a/h1n1. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*. 2011; 9(2):89–115.
38. Koopman J. Modeling infection transmission. *Annual Review of Public Health*. 2004; 25:303–326.
39. Meza R, Haaf K, Kong CY, Erdogan A, Black WC, Tammemagi MC, Choi SE, Jeon J, Han SS, Munshi V, et al. Comparative analysis of 5 lung cancer natural history and screening models that reproduce outcomes of the nlst and plco trials. *Cancer*. 2014; 120(11):1713–1724. [PubMed: 24577803]
40. Walter E, Lecourtier Y, Happel J. On the structural output distinguishability of parametric models, and its relations with structural identifiability. *Automatic Control, IEEE Transactions on*. 1984; 29(1):56–57.
41. Fung IC-H. Cholera transmission dynamic models for public health practitioners. *Emerging themes in epidemiology*. 2014; 11(1):1. [PubMed: 24520853]
42. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A*. 1927; 115:700–721.
43. Dunworth, JB. Master's thesis. Ohio State University; 2011. Nonlinear incidence of waterborne diseases.
44. Haas, CN., Rose, JB., Gerba, CP. Quantitative microbial risk assessment. John Wiley & Sons; 1999.
45. Wearing HJ, Rohani P, Keeling MJ. Appropriate models for the management of infectious diseases. *PLoS Medicine*. 2005; 7(7)
46. Lloyd AL. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. *Theoretical Population Biology*. 2001; 60:59–71. [PubMed: 11589638]
47. Hethcote HW. Simulations of pertussis epidemiology in the united states: effects of adult booster vaccinations. *Mathematical biosciences*. 1999; 158(1):47–73. [PubMed: 10209936]
48. Mosley WH, Benenson AS, Barui R. A serological survey for cholera antibodies in rural east pakistan: 2. a comparison of antibody titres in the immunized and control populations of a cholera-

- vaccine field-trial area and the relation of antibody titre to cholera case rate*. Bulletin of the World Health Organization. 1968; 38(3):335. [PubMed: 5302923]
49. CIA. Central Intelligence Agency World Factbook - Haiti. 2011. <https://www.cia.gov/library/publications/the-world-factbook/geos/ha.html>
 50. Dunworth, JB. PhD thesis. The Ohio State University; 2011. Nonlinear Incidence of Waterborne Diseases.
 51. CIA. Central Intelligence Agency World Factbook - Angola. 2015. <https://www.cia.gov/library/publications/the-world-factbook/geos/ao.html>
 52. Weil AA, Khan AI, Chowdhury F, LaRocque RC, Faruque ASG, Ryan ET, Calderwood SB, Qadri F, Harris JB. Clinical outcomes of household contacts of patients with cholera in Bangladesh. Clinical Infectious Diseases. 2009; 49:1473–1479. [PubMed: 19842974]
 53. Cobelli C, DiStefano JJ. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. American Journal of Physiology - Regulatory, Integrative and Comparative Physiology. 1980; 239(1):R7–R24.
 54. Rothenberg T. Identification in parametric models. Econometrica. 1971; 577
 55. Jacquez JA, Greif P. Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. Mathematical Biosciences. 1985; 77(1–2):201–227.
 56. Komorowski M, Costa MJ, Rand DA, Stumpf MPH. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. Proceedings of the National Academy of Sciences. 2011; 108(21):8645–8650.
 57. Cintron-Arias A, Banks H, Capaldi A, Lloyd A. A sensitivity matrix methodology for inverse problem formulation. Journal of Inverse and Ill-posed Problems. 2009; 17:1–20.
 58. Eisenberg MC, Hayashi MAL. Determining identifiable parameter combinations using subset profiling. Mathematical Biosciences. 2014; 256:116–126. [PubMed: 25173434]
 59. Akaike H. Factor analysis and aic. Psychometrika. 1987; 52(3):317–332.
 60. Sack DA, Bradley Sack R, Chaignat Claire-Lise. Getting serious about cholera. New England Journal of Medicine. 2006; 355(7):649–651. [PubMed: 16914700]
 61. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmuller U, Timmer J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009; 25(15):1923–9. [PubMed: 19505944]
 62. Butler D. Models overestimate ebola cases. Nature. 2014; 515(18)
 63. Rivers C. Ebola: models do more than forecast. Nature. 2014; 515(492)
 64. Weitz JS, Dushoff J. Modeling post-death transmission of ebola: Challenges for inference and opportunities for control. Scientific reports. 2015; 5
 65. Eisenberg MC, Eisenberg JN, D'Silva JP, Wells EV, Cherng S, Kao Y-H, Meza R. Forecasting and uncertainty in modeling the 2014–2015 ebola epidemic in west africa. 2015 arXiv preprint, pp. arXiv:1501.05555 [q-bio.PE].
 66. Pascual M, Rodo X, Ellner SP, Colwell R, Bouma MJ. Cholera dynamics and El Niño-Southern oscillation. Science. 2000; 289:1766–1769. [PubMed: 10976073]
 67. Pascual M, Chaves LF, Cash B, Rodo X, Yunus M. Predicting endemic cholera: the role of climate variability and disease dynamics. Climate Research. 2008; 36:131–140.
 68. Reiner RC, King AA, Emch M, Yunus M, Faruque ASG, Pascual M. Highly localized sensitivity to climate forcing drives endemic cholera in a megacity. Proceedings of the National Academy of Sciences. 2012; 109(6):2033–2036.
 69. Alam A, LaRocque RC, Harris JB, Vanderspurt C, Ryan ET, QF, Calderwood SB. Hyperinfectivity of human-passaged *Vibrio cholerae* by growth in the infant mouse. Infection and Immunity. 2005; 73(10):6674–6679. [PubMed: 16177344]
 70. Bertuzzo E, Casagrandi R, Gatto M, Rodriguez-Iturbe I, Rinaldo A. On spatially explicit models of cholera epidemics. Journal of the Royal Society Interface. 2010; 7(43):321–333.
 71. Eisenberg MC, Shuai Z, Tien JH, van den Driessche P. A cholera model in a patchy environment with water and human movement. Mathematical Biosciences. 2013; 246(1):105–112. [PubMed: 23958383]

72. Righetto L, Bertuzzo E, Casagrandi R, Gatto M, Rodriguez-Iturbe I, Rinaldo A. Modelling human movement in cholera spreading along fluvial systems. *Ecohydrology*. 2011; 4(1):49–55.
73. Alexanderian A, Gobbert MK, Fister KR, Gaff H, Lenhart S, Schaefer E. An age-structured model for the spread of epidemic cholera: Analysis and simulation. *Nonlinear Analysis: Real World Applications*. 2011; 12(6):3483–3498.
74. Galvani AP, Reluga TC, Chapman GB. Long-standing influenza vaccination policy is in accord with individual self-interest but not with the utilitarian optimum. *Proceedings of the National Academy of Sciences*. 2007; 104(13):5692–5697.
75. Merl D, Johnson LR, Gramacy RB, Mangel M. A statistical framework for the adaptive management of epidemiological interventions. *PloS One*. 2009; 4(6):e5807. [PubMed: 19503812]
76. Yaesoubi R, Cohen T. Dynamic health policies for controlling the spread of emerging infections: influenza as an example. *PloS one*. 2011; 6(9):e24043. [PubMed: 21915279]
77. Fenichel EP, Castillo-Chavez C, Ceddia M, Chowell G, PA, Parra G, Hickling GJ, Holloway G, Horan R, Morin B, Perrings C, et al. Adaptive human behavior in epidemiological models. *Proceedings of the National Academy of Sciences*. 2011; 108(15):6306–6311.
78. Shea K, Tildesley MJ, Runge MC, Fonnesbeck CJ, Ferrari MJ. Adaptive management and the value of information: learning via intervention in epidemiology. *PLoS Biol*. 2014; 12(10):e1001970. [PubMed: 25333371]
79. Moore SM, Lessler J. Optimal allocation of the limited oral cholera vaccine supply between endemic and epidemic settings. *Journal of The Royal Society Interface*. 2015; 12(111):20150703.
80. Hayashi MA, Eisenberg MC. Effects of adaptive protective behavior on the dynamics of sexually transmitted infections. *Journal of theoretical biology*. 2016; 388:119–130. [PubMed: 26362102]
81. Feachem, R., Bradley, D., Garelick, H., Mara, D. *Vibrio cholerae* and cholera. In: Feachem, RG., editor. Sanitation and disease – health aspects of excreta and wastewater management. John Wiley & Sons; 1983. p. 297–325.
82. Hood MA, Ness GE, Rodrick GE. Isolation of *Vibrio cholerae* serotype O1 from the eastern oyster, *Crassostrea virginica*. *Appl Environ Microbiol*. 1981; 41(2):559–560. [PubMed: 7235700]

Figure 1. Diagrams of study models

First row (left to right): Exponential model, Dose Response model, Asymptomatic model. Second row (left to right): Gamma model and Progressive Susceptibility model. Red compartments represent the infected population and red arrows represent person-person transmission. Blue compartments represent pathogen concentration in water while blue arrows represent pathogen-person transmission. Black compartments are susceptible or partially susceptible, while white compartments are immune. Grey arrows indicate pathogen shedding.

Figure 2. Fits to 100-day simulated model data without noise (left column), with normal noise (middle column), and with Poisson noise (right column), using naive starting parameters. Each row represents data simulated by the model labeled at right. Model fits are overlaid, thus obscuring some of the model fits in the figure. All fitting models were able to capture the mean epidemic data despite added noise and model misspecification.

Figure 3.

Percent deviation of estimates from true parameter values, grouped by parameter for A) the simulated 100-day data and B) the simulated 3-year data. The model used to fit the data and estimate the parameter is indicated by color. The median across all estimates (i.e., across added noise type, simulation model, and data duration) is marked with a black point in the distribution and the black dashed line represents $\pm 20\%$ deviation. Distribution ranges for the β_W , α , and ξ subplots are trimmed for visibility. C) Scatterplot of β_W and ξ estimates by colored fitting model. Data points in the distribution tails were truncated for ease of viewing, but see Supplementary Figure S2 for the full plot.

Figure 4. Forecasts from informed starting values to 100-day data (indicated by row) up through 10 days, (left column), 30 days (middle column), and 50 days (right column) with added normal noise. Model fits are overlaid, thus obscuring some of the model fits in the figure.

Figure 5. Parameter estimates derived from naive starting parameter forecasts with 10, 30, and 50 days of data, organized A) by parameter and B) by added noise.

Figure 6.
 “Naive” (left) and “informed” (right) starting parameter fits to 2006 Angola epidemic data.

Model parameters. Additional details about the baseline values for each parameter are given in Section 2.2 and the Supplementary Information (SI).

Table 1

Parameter	Description	Value	Units	Source	Models
μ	natural birth/death rate in Angola	$\frac{1}{(55 \times 365)}$	$days^{-1}$	[51]	All
γ	1/(mean duration of infectiousness), or recovery rate	$\frac{1}{4}$	$days^{-1}$	[2, 14, 20, 52]	All
β_I	person-person transmission rate	0.25	$days^{-1}$	[8, 14, 20]	All except Asymptomatic
β_{IA}	person-person transmission rate in asymptomatics	$\frac{5}{32}$	$days^{-1}$	See Model Description, Section 2.2	Asymptomatic
β_{IS}	person-person transmission rate in symptomatics	$\frac{5}{16}$	$days^{-1}$	See Model Description, Section 2.2	Asymptomatic
β_W	water-borne transmission rate	0.50	$days^{-1}$	[8, 14, 20]	All
α	loss of immunity rate	$\frac{1}{365}$	$days^{-1}$	[6, 27, 28], See Section 2.2	All except Asymptomatic
α_{IA}	loss of immunity rate in asymptomatics	$\frac{1}{324.4}$	$days^{-1}$	See Section 2.2	Asymptomatic
α_{IS}	loss of immunity rate in symptomatics	$\frac{1}{730}$	$days^{-1}$	See Section 2.2	Asymptomatic
ξ	decay rate of cholera in water	$\frac{1}{100}$	$days^{-1}$	[14, 20]	All
k	1/(population at risk · reporting rate)	$\frac{1}{50,000}$	$people^{-1}$	[20], See Section 2.2, SI	All except Asymptomatic
k	proportion symptomatic cases/(population at risk · reporting rate)	$\frac{0.2}{50,000}$	$people^{-1}$	[20], See Section 2.2, SI	Asymptomatic
K	bacteria concentration that gives 50% chance of infection	0.4	$unitless$	[50], See Model Description, Section 2.2	Dose Response
u_1	modifies the susceptibility of individuals to direct infection	10	$unitless$	See Model Description	Progressive Susceptibility
u_2	modifies the susceptibility of individuals to waterborne infection	10	$unitless$	See Model Description	Progressive Susceptibility

Estimated % CV for each parameter. A % CV > 100 means that the parameter is practically unidentifiable. A ‘-’ means that the parameter is not applicable to the model.

Table 2

Parameters	Exponential	Dose Response	Asymptomatic	Gamma	Progressive
β_I	0.016	0.019	-	0.011	0.016
β_{IS}	-	-	4.63×10^5	-	-
β_{IA}	-	-	4.63×10^5	-	-
β_W	0.56	0.27	0.82	0.34	0.48
α	0.34	0.094	-	0.11	0.23
α_S	-	-	2.23	-	-
α_A	-	-	0.52	-	-
ξ	0.66	0.33	0.96	0.37	0.55
k	0.025	0.0081	0.040	0.042	0.044
u_1	-	-	-	-	0.55
u_2	-	-	-	-	8.33

Table 3

Parameter estimates and % CVs (shown in parentheses) for the 2006 Angola epidemic data.

(a) <i>Informed</i> starting parameters					
Parameters/AIC	Exponential	Dose Response	Asymptomatic	Gamma	Progressive
β_i	0.146 (14.02)	0.1981 (3.58)	0.1474 (16.56)	0.2631 (0.82)	0.034 (131.43)
β_w	0.6033 (14.23)	0.5055 (18.47)	0.6089 (17.35)	1.2563 (49.59)	1.0985 (24.61)
ξ	0.0544 (35.62)	0.0148 (25.76)	0.0531 (43.13)	0.0074 (53.79)	0.047 (53.54)
α	0.0051 (27.68)	0.002 (7.48)	0.005 (52.75)	0.0013 (14083)	0.0007 (648.07)
k	1.3661e-05 (1.02)	1.3021e-05 (0.69)	1.3653e-05 (1.10)	1.1255e-05 (1.40)	1.7878e-05 (1.94)
AIC	3667	3679	3672	4305	3038
θ_{c_0}	2.9967	5.8467	3.0243	6.0768	4.529

(b) <i>Naïve</i> starting parameters					
Parameters/AIC	Exponential	Dose Response	Asymptomatic	Gamma	Progressive
β_i	0.1687 (7.20)	0.1804 (6.32)	0.1435 (11.05)	0.2521 (1.27)	7.58e-8 (5.56e7)
β_w	0.7222 (14.04)	0.4419 (20.09)	0.7003 (11.60)	1.5945 (62.94)	0.9638 (6.66)
ξ	0.0357 (26.99)	0.0197 (31.37)	0.045 (31.13)	0.0066 (68.81)	0.0675 (18.81)
α	0.0037 (19.47)	0.0023 (11.69)	0.0049 (26.76)	0.0053 (2.48)	0.0003 (304.27)
k	1.3632e-05 (0.92)	1.3171e-05 (0.69)	1.4147e-05 (1.25)	1.1643e-05 (1.29)	1.953e-05 (6.58)
AIC	3705	3644	3671	4158	2876
θ_{c_0}	3.5628	5.1392	3.3741	7.3849	3.8546