Fitting outbreak models to data from many small norovirus outbreaks

Eamon B. O'Dea
University of Texas at Austin

Kim M. Pepin
Colorado State University, kim.m.pepin@aphis.usda.gov

Ben A. Lopman
Health Protection Agency

Claus O. Wilke
University of Texas at Austin
Fitting outbreak models to data from many small norovirus outbreaks

Eamon B. O’Dea, Kim M. Pepin, Ben A. Lopman, Claus O. Wilke

Abstract
Infectious disease often occurs in small, independent outbreaks in populations with varying characteristics. Each outbreak by itself may provide too little information for accurate estimation of epidemic model parameters. Here we show that using standard stochastic epidemic models for each outbreak and allowing parameters to vary between outbreaks according to a linear predictor leads to a generalized linear model that accurately estimates parameters from many small and diverse outbreaks. By estimating initial growth rates in addition to transmission rates, we are able to characterize variation in numbers of initially susceptible individuals or contact patterns between outbreaks. With simulation, we find that the estimates are fairly robust to the data being collected at discrete intervals and imputation of about half of all infectious periods. We apply the method by fitting data from 75 norovirus outbreaks in healthcare settings. Our baseline regression estimates are 0.0037 transmissions per infective-susceptible-day, an initial growth rate of 0.27 transmissions per infective-day, and a symptomatic period of 3.35 days. Outbreaks in long-term-care facilities had significantly higher transmission and initial growth rates than outbreaks in hospitals.

Introduction
A common and difficult problem in epidemiology is to estimate rates of disease spread. Accurate estimates of these and other population parameters are crucial in the evaluation of disease control measures (Anderson and May, 1992; Keeling, 2005; Halloran et al., 2009) or biological hypotheses (Lively, 2010). Heterogeneity complicates the problem of obtaining such estimates. For example, a person’s risk of infection depends on contact rates and acquired immunity, and these quantities can vary widely between people and outbreaks.

Norovirus (NoV) epidemiology provides a fine case in point of the need for models to accommodate heterogeneity. Noroviruses are the most common cause of diarrheal disease in the United States, causing an estimated 21 million cases (Scallan et al., 2011) and 71,000 hospitalizations per year (Lopman et al., 2011). A genetically diverse group of strains is often circulating within a population. New strains of the predominant genogroup 2 genotype 4 (GI.4) taxon appear regularly over time (Glass et al., 2009), and a person’s risk of infection, given exposure, likely depends on both the antigenicity of the virus and the type-specific immunity developed from the person’s previous exposure (Cannon et al., 2009). Other important heterogeneities include innate susceptibility (which depends on a person’s histo-blood group antigens and secretor status) and age-specific risks of exposure. Outbreak investigations (Evans et al., 2002; Thornley et al., 2011; Wikswo et al., 2011) have provided convincing evidence that single vomiting incidents in crowded settings can lead to scores of secondary cases. Models that account for both between-individual and between-population heterogeneity are needed to obtain the accurate parameter estimates required for predicting outbreak dynamics and implementing effective controls. At present, control measures are based on general infection-control principles (Centers for Disease Control and Prevention, 2011) and thus are likely to be somewhat inefficient.

A further complication for modeling norovirus transmission is that it often occurs in small outbreaks. The transmission and
recovery times of cases in small outbreaks are correlated (Rida, 1991), which makes estimation difficult when using data from a single outbreak. An obvious solution to reducing the inaccuracy caused by within-outbreak correlations in data is to base estimates on data from multiple outbreaks.

Methods for estimating parameters from multiple outbreaks have been described before, but often have been developed for smaller data sets and computing resources than what are now available. For example, the previous approaches of Becker (1979) and Becker (1991) assumed only the observation of the final state of each outbreak was available, used moments estimators, did not formulate general a regression model to allow for variation in parameters between outbreaks, and may be implemented with pencil and paper. Our norovirus outbreak data set includes the full observation of a large number of outbreaks and a number of covariates that are likely to affect parameters. We thus here employ a different method that operates on the full observation of outbreaks, uses maximum-likelihood estimators, models the effect of covariates on outbreak parameters within a general regression framework, and exploits modern computing power to find estimates and their confidence intervals.

We propose a general approach to fitting data from many small outbreaks. Using simulated data, we assessed the performance of the proposed method as a function of the number of outbreaks in the data, the rounding of measurements to regular intervals of observation, the number of missing observations, and the imputation of missing observations. When the number of outbreaks was large, we found the performance to be satisfactory for data sets with realistic levels of all of these challenging features. Fitting our model to data from a large number of real norovirus outbreaks in health-care facilities, we found a distinct increase in transmission and initial growth rates in long-term-care facilities relative to hospitals. We examined the fit of the model and found the most noticeable defect to be lower-than-observed prediction of the initial growth of the outbreaks. However, the predicted dynamics became more accurate over time such that predictions never deviated widely from observations.

Methods

We developed the methods described in this section to fit a model of the outbreak dynamics of norovirus based on data from a large survey of gastroenteritis in health-care facilities in the former County of Avon, England. In this study, the events of symptom onset and recovery were recorded on a daily basis for cases of gastroenteritis in both care staff and patients in 15 hospitals and 135 long-term-care facilities over a year-long period in 2002–2003, and these events were classified into a total of 271 separate outbreaks (Lopman et al., 2004b). These outbreaks were for the most part small; the range in total cases spans from 2 to 90 cases and the median is 13 cases.

We begin by presenting our estimation methods. With the method defined, we then describe assumptions and imputation procedures used to prepare our data for application of the method. To complete the model specification for our application, we next describe the variables of the data chosen to be predictors of how parameters vary among outbreaks. Finally, we provide details about methods of simulation, calculation of confidence intervals, and choice of software.

Model

Although our aim is to introduce a general approach, we aim to do so by way of example. Thus we describe our methods in terms of a specific model choice made for the norovirus data. However, we do provide references to relevant results in the regression literature to indicate the full scope of this approach.

The states and transition rules for the model we adopt for individual outbreaks are as follows. The population consists of a fixed number of people of one or more types. The term type here identifies people by the rules governing their movement between different states with respect to norovirus infection. At the beginning of an outbreak, there is some positive number of people in an exposed, or latent, state for at least one of the types. This state represents people who have been exposed to an infection source and have a latent infection but are not contagious. They move to an infective state after an incubation period of fixed duration. The infective state represents contagious people, and for simplicity we assume that all contagious people are symptomatic. A susceptible state represents people who are susceptible to infection. Thus each susceptible of type i moves to the latent state at the first point of a Poisson process with rate $\beta_i(t)$, where $\beta_i$ is the transmission rate for type- i susceptibles and $\gamma(t)$ is the number of infectives at time t. All infective types have the same level of contagiousness and have gamma-distributed symptomatic periods with the same dispersion parameter, but the mean symptomatic period may differ between types. Further, types that represent care staff are moved into an infective-but-removed state when the time they have spent in the infective state exceeds a threshold of fixed duration. This transition rule represents the effect of infection-control policies that prevent staff from working contagious. At the end of their symptomatic periods, infective and infective-but-removed people are moved into a recovered state. The recovered state represents individuals that gain immunity over the course of the outbreak. The outbreak ends when the number of infected people reaches zero. In summary, our outbreak model is the widely studied susceptible-exposed-infective-recovered (SEIR) model with four customizations for our application. First, we allow people to vary in susceptibility and expected duration of infectiousness. Second, we do not make our transmission rate depend on the total number of people in the population. This departure prevents the need for the total number of people to be estimated, and it is appropriate in small populations when an infective person may be able to infect every susceptible person in the population with approximately the same probability. For example, Forrester and Pettitt (2005) did not find that inclusion of the total population size significantly improved the fit of a model of methicillin-resistant Staphylococcus aureus (MRSA) outbreaks within an intensive-care unit. Third, we do not assume that latent periods and infectious periods are exponentially distributed. Our approach is more realistic because it allows the probability of a person leaving a latent or infectious state to depend on how long she has been in that state. Fourth, we shunt some of the infectives into an infective-but-removed state to represent the isolation of contagious staff from the population.

As indicated in our outbreak model description, the rate at which a susceptible acquires infection from an infective may vary among members of a population, and we use the word type in a general sense to refer to subsets of the population that are assumed to be the same with respect to such variation. With multiple-outbreak data, we further define types as unique to individual outbreaks. In other words, we make no general assumption that people in different outbreaks may be modeled with the same parameters. We shall later choose a particular linear model that controls the extent to which parameters may vary among types, but many other choices for such models are possible within this framework. Types thus represent the fundamental unit of variation in this framework, and the likelihood function naturally breaks apart into factors for each type.

For each type, the recovery-time and transmission-time parts of the likelihoods further factor apart into common density functions. The simplicity of these functions belies an involved construction,
available in Kalbfleisch and Prentice (2002), as the product integral of the likelihood of events in infinitesimal time steps, where the likelihood of each time step is conditional on the history of the model up until that step time. We shall introduce the full likelihood by introducing each of these functions in turn.

For type-i people, the recovery-time part of the likelihood is

$$l_{rec}(\mu_i, \rho) = \prod_{j=1}^{k_i} \frac{1}{\Gamma(1/\rho)(\rho \mu_i)^{1/\rho}} l_{ij}^{1/\rho - 1} \exp \left( - \frac{l_{ij}}{\rho \mu_i} \right),$$

where $k_i$ is the number of type-i people infected over the course of an outbreak, $l_{ij}$ denotes the length of the symptomatic period of the jth type-i infection, $\mu_i$ is the mean of the symptomatic period of type-i infections, and $\rho$ is the dispersion parameter, which we take to be the same for all types of infections. Eq. (1) represents the likelihood function for a joint distribution of gamma-distributed random variables. Recall that per our model definition, the symptomatic periods $l_{ij}$ are gamma distributed.

The transmission-time part of the likelihood for type-i people is

$$l_t(\beta_i, X_i(0)) = X_i(0)^{1/2}(X_i(0) - k_i) \exp \left[ - \beta_i \tau_i(X_i(0) - k_i) \right] \prod_{j=1}^{k_i} \beta_j Y_{ij} \exp \left( - \beta_j h_{ij} \right),$$

where $X_i(0)$ is the number of initial susceptibles, $\tau_i$ is the cumulative exposure of such people at the end of an outbreak (i.e., the total area under $\gamma(t)$), $Y_{ij}$ is the number of infectives present when the jth such person becomes infected, $h_{ij}$ is the cumulative exposure of the jth such person when infected. Further discussion of this likelihood function is provided in the Appendix.

In many cases, converting the data to a minimally sufficient form may be desirable for the preservation of patient privacy. An example of such a form would be to summarize the data as $\sum h_{ij}$, $\sum h_{ij}$, $k_i$, and $\tau_i$ for each type i. Such a form would still allow for calculation of maximum-likelihood estimates and Hessian-based (Wald) confidence intervals.

This minimally sufficient form of the data also illustrates the robustness of estimates to some imperfections of the data. The $h_{ij}$ and $l_{ij}$ only affect the likelihood through the sums $\sum h_{ij}$ and $\sum l_{ij}$. Thus some error in our calculation of $h_{ij}$ and $l_{ij}$ should not bias our estimates too much as long as the average error is close to zero, and thus great certainty about $\gamma(t)$ is not necessary. For example, if we underestimated $\gamma(t)$ at some points as a result of asymptomatic infectives being present in reality and we overestimated $\gamma(t)$ at some other points as a result of misdiagnoses, those errors may cancel each other to some extent. Also, sometimes the data consist of only the times at which people stop being infectious—for example, when people are isolated after being identified as infectious. In such cases total exposure could still be estimated by using a kernel-smoothing method (Lau and Yip, 2008).

The likelihood (2) can be parameterized differently as

$$l_t(\beta_i, r_i) = (r_i/\beta_i)^{1/2}(r_i - k_i) \exp [r_i(\beta_i k_i - r_i)] \prod_{j=1}^{k_i} \beta_j Y_{ij} \exp ( - \beta_j h_{ij}),$$

where $r_i = \beta_i X_i(0)$ is the initial per-infective incidence rate. In our application, we choose to estimate $r_i$ instead of $X_i(0)$ because $r_i$ is easier to interpret in the context of our data. For brevity, we refer to $r_i$ as the initial growth rate.

The full likelihood function that we use for an $n$-outbreak data set is then

$$l(\beta, r, \mu, \rho) = \prod_i l_t(\beta_i, r_i) l_{rec}(\mu_i, \rho),$$

where we use boldface to denote vectors with elements equal to the parameters for each type $i$.

To make use of previous results from statistical theory as well as to use conventional language when writing about our model, we shall next present our model as a generalized linear model (GLM). GLMs are a broad class of statistical models that includes many commonly used regression models. A GLM consists of three components: (i) a density function from the exponential family, (ii) a linear model that maps predictive variables to a predictor, and (iii) a link function that maps the predictor to the mean of the density function.

Our likelihood functions, (1) and (3), fit the definition of exponential family densities. That is not to say that the transmission and recovery times from a small outbreak are independent random variables with those densities. In fact, they may be highly correlated (Rida, 1991). But the situation is analogous to that of GLMs for longitudinal data, where ignoring within-subject correlations increases the variance of estimates but still leads to accurate estimates in the limit of data from a large number of independent subjects (Liang and Zeger, 1986).

We obtain a linear model by associating each type of person in the model with a set of predictive variables. In the application to norovirus we describe here, such predictive variables are, for example, the type of facility in which an outbreak occurred (e.g., hospital or long-term-care facility). We combine these predictive variables into a design matrix $Z$, which has a row for each type $i$ and a column for each predictive variable. The linear mapping from multiple predictive variables to a linear predictor is achieved by multiplying the design matrix with a vector of regression parameters $\beta$.

As link function, we chose the natural log, which tended to perform better than other potential link functions in our application. For example, for transmission-rate estimates $\beta_i$, we let $\log \beta_i = Z_i \cdot \mathbf{c}_i$, where $Z_i$ is row $i$ of the design matrix and $\mathbf{c}_i$ are our regression parameters for the transmission rates.

The conditions for consistency and asymptotic normality of parameter estimates for GLMs have been given by Fahrmeir (1985). For outbreak data, the values of predictive variables in the data are likely to be somewhat randomly determined, in which case the conditions given by Ding and Chen (2006a,b) apply.

One standard condition for consistency is that the true value of the parameter does not lie on the boundary of parameter space. That condition would seem to be violated for data sets in which the number of cases in an outbreak $k_i$ is equal to the number of susceptibles $X_i(0)$ because the transmission-time part of the likelihood, (2), is defined only when $X_i(0) \geq k_i$. However, given that we approximate the discrete quantity $X_i(0)$ with a continuous one for the purpose of fitting the model, it seems reasonable to consider $k_i - 0.5$ as the lower bound of $X_i(0)$ and to say that

$$l_t(\beta_i, X_i(0)) = X_i(0)^{1/2}(X_i(0) - k_i) \exp (0) \prod_{j=1}^{k_i} \beta_j Y_{ij} \exp ( - \beta_j h_{ij}),$$

when $X_i(0)$ is in $(k_i - 0.5, k_i]$. Then the true value of $X_i(0)$ is guaranteed not to be on the boundary and standard consistency results apply. In the Appendix, we provide an alternative proof of consistency for our model in the simple case that all outbreaks share the same parameters.

Evidence that the model performs well in realistic situations appears in the Results section. We were able to recover from
simulated data the parameters for the non-trivial model that we fitted in our application.

We estimated the transmission rate and initial growth rate by maximizing the transmission-time factors in (4) given the outbreak data, using the Newton–Raphson method as implemented in the AD Model Builder (Fournier et al., 2011). To keep the Newton–Raphson search for maximum-likelihood estimates in the feasible parameter space, we added a penalty to the log likelihood whenever the implied final number of susceptibles \( X = X(0) - k_i \) for an outbreak was too close to zero, \( x < e \). The penalty was of the form \( C(x - e) \), where \( C \) is an arbitrary numeric constant which we set to \( C = 0.01 \). Likewise, whenever \( x < e \), we replaced \( x \) by \( e/2 \). Throughout this work, we used \( e = 0.001 \).

We estimated the mean infectious period and the dispersion of the infectious period by using the glm function in R (R Development Core Team, 2010). By default, the dispersion parameter for gamma GLMs is estimated via the moments estimate for the coefficient of variation, perhaps because estimates based on the residual deviance are sensitive to small values in the data (Venable and Ripley, 2002, p. 9). Consistently, we typically found default estimates to be more accurate than ML estimates when fitting small, simulated data sets and we found them to be very similar to ML estimates when fitting large data sets. To be consistent in our treatment of both small and large data sets, we used the default estimate of the dispersion parameter throughout this work.

Data

The norovirus (NoV) data we analyze here originated in a prospective surveillance program in hospitals and long-term-care facilities in England (Lopman et al., 2004a,b). We analyzed the dynamics of 75 outbreaks laboratory-confirmed to be caused by NoV in which a total of 1523 cases of gastroenteritis occurred among patients and staff. We selected these data from the larger data set produced by the surveillance program as follows.

Most records of infections that were attributed in whole or in part to norovirus included the dates of both the onset of and the recovery from symptoms. However, in many records both dates were missing, and in most outbreaks some records lacked at least one date.

We discarded all records from outbreaks in which more than 55% of the dates of recovery were missing. In the remaining outbreaks, we replaced missing dates of recovery with the corresponding onset date plus the median symptomatic period from complete records in that outbreak. These replacements were done as a preparation for the estimation of the transmission rates and were not included when estimating symptomatic periods.

We discarded all records where the onset date was missing. This practice is unlikely to introduce a large bias as long as a relatively small number of onset dates are discarded. We made sure that this number was relatively small by using data only from outbreaks in which the number of records that were missing onset dates was less than 7% of the number of records that were not missing onset dates.

We made several simplifying assumptions. We assumed a person is infective only when symptomatic, which is supported by Sukhrie et al. (2012). We further assumed that staff move to the infective-but-removed state after one day of symptoms, in accordance with an infection control policy. Of course, staff with norovirus symptoms are likely to stop working sooner than that in many cases. But it seems likely that all infective staff have some small probability of transmitting the disease to others before they leave. Zelnner et al. (2013) found that household outbreak data supported a model in which infectiousness is highest at the onset of symptoms, which is often when vomiting occurs. A one-day infectious period is a simple way of modeling the effect of people who may, in fact, only be present for shorter periods with higher infectiousness.

We also assumed that the latent period is fixed at 24 h, which falls well within the reported range of 12–48 h (Centers for Disease Control and Prevention, 2011). We keep the transmission-time part of the likelihood positive, we assumed a small, background hazard of infection \( 10^{-8} \) that of an infective) triggered illness in cases when no infectives were present. We also assumed that the number of initial infective people was equal to the number of people reporting symptoms on the first day of the outbreak. Finally, we assumed that any changes in state happen at the same time each day.

### Predictive variables

The predictive variables that determined our design matrices were as follows. The data were collected over the course of a one-year period beginning in April 2002, and we categorized the data into two groups by the period in which they began: spring–summer refers to outbreaks that started between April 1 and October 1 of the study year; fall–winter refers to outbreaks that began in the remainder of the study year. The period variable allows for variation in transmission rate as a result of seasonality of NoV.

As an additional predictive variable, we include what type of facility the outbreak occurred in, hospital or long-term-care facility (LTCF).

The third predictive variable we use is size class. We classify units in which the number of beds is less than or equal to the median number of beds as small. We classify the other units as large. This classification was done separately for hospital and LTCF units because LTCF units are usually larger than hospital units. For the hospitals, the small units have 6–22 beds and the large units have 24–33 beds. For the LTCF units, the small units have 6–34 beds and the large units have 36–66 beds. The size class variable allows the number of initial susceptibles to depend on the approximate total number of people in each unit. The variable also allows population sizes to affect contact rates.

The fourth predictive variable we use is case type, the two types being patient and staff. Case type is the only predictive variable that varied within outbreaks.

We use a facility–size-class–period–case-type combination with a relatively large amount of data as the reference group. Specifically, the reference group comprises outbreaks that occurred among patients in large care-units of hospitals that began between October 2002 and April 2003. The estimated rate parameter for the reference group serves as the coefficient of the intercept of the linear model. Estimates for other coefficients then inform us of how moving away from the reference group changes rate estimates. Table 1 contains the distribution of outbreaks among the levels of the predictive variables.

### Table 1

<table>
<thead>
<tr>
<th>Facility</th>
<th>Season</th>
<th>Size class</th>
<th># outbreaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTCF</td>
<td>Fall–winter</td>
<td>Small</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Spring–summer</td>
<td>Small</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>4</td>
</tr>
<tr>
<td>Hospital</td>
<td>Fall–winter</td>
<td>Small</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Spring–summer</td>
<td>Small</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>11</td>
</tr>
</tbody>
</table>
predictors for all our parameters is: \( 1 + \text{isLTDF} + \text{isSmall} + \text{isSpringSummer} + \text{isStaff} \).

**Confidence intervals**

To obtain confidence intervals for the estimates, we performed a parametric bootstrap. Data were simulated according to our outbreak model with the estimated parameters. Each simulation produced data from a set of outbreaks equal in size to the set that we fit, with each outbreak in the simulation matching an outbreak in the fitted data in terms of initial number of infectives, predictive variables, fraction of case records with missing onset and recovery times, and fraction of cases with missing recovery times. Percentile confidence intervals for regression coefficients were estimated from 10,000 simulation replicates.

**Simulation**

We used simulation to investigate how the bias and variance of our estimates depend on the number of outbreaks that they are based on as well as the amount of missing information. We also used simulation to generate bootstrap confidence intervals.

Simulations began with some initial numbers \( X_i^{(0)} \) of type-\( i \) susceptibles of one or more types. To initiate the outbreak, some additional susceptibles were added and moved into a latent state. All people entering the latent state moved to the infective state after a fixed time period. Type-\( i \) people entering the infective state moved on to the recovered state after a gamma-distributed time period with mean \( \mu_i \) and dispersion parameter \( \rho \). For types of infectives representing care staff, movement to the infective-but-removed state preceded movement to the recovered state if and when time spent in the infective state exceeded a predetermined threshold. Every time the number of infectives or susceptibles changed, the time of a potential transmission event was calculated by drawing from an exponential distribution with rate \( Y \sum \beta_i X_i \), where \( Y \) is the number of (non-removed) infectives and \( \beta_i \) is the transmission rate for susceptibles of type \( i \). If the potential transmission was sooner than the next change in \( Y \), a type of susceptible was chosen with probability proportional to \( \beta_i X_i \) and moved into the latent state. Simulations stopped when the number of latent, infective, and infective-but-removed people reached zero. The output of the simulations was a record for each person infected giving transition times.

Our simulation experiment had a full factorial design, with the number of outbreaks \( n \) being 1, 10, or 100; the fraction of recovery times imputed being either zero or approximately the highest such fraction in our real data (0.53); the fraction of records missing both onset and recovery times being either zero or approximately the highest such fraction in our real data (0.05); and onset and recovery times being either rounded to days or exact. For each combination of factor levels, we simulated data and attempted to fit it 10,000 times. These simulations had just one type, just one initially infected person, a transmission rate \( \beta \) of 0.0037, an initial growth rate \( r \) of 0.2664, a latent period of 24h, and infectious periods with a mean \( \mu \) of 3.32 days and a dispersion parameter \( \rho \) of 0.58, and no threshold time beyond which the infectives were moved into an infective-but-removed state.

The details of the simulation of missing data were as follows. First, an outbreak was simulated as usual. Second, the number of case records to remove was calculated as the largest integer less than the total number of case records times the parameter value for the fraction of records missing onset and recovery times. This number of records were selected at random and removed. Then the number of case records for which to impute recovery times was calculated as the largest integer less than the total number of remaining case records times the parameter value for the fraction of records imputed. This number of records were selected at random and given imputed recovery times.

Once-daily observation of the outbreak was simulated by rounding transition times down to the nearest whole day. Outbreaks were started at random times in the first day to prevent the rounding from having artificial effects on the data from small outbreaks.

Our gradient-based optimization code for model fitting, which worked well at estimating transmission rate parameters when the number of outbreaks was large, did not work well when the number of outbreaks was small. So we used specialized code to fit the models of the simulation study, which were more analytically tractable by virtue of not having linear predictors. The Appendix describes the basis for this code, which always finds the maximum-likelihood estimate if it exists and identifies cases in which no such estimate exists.

**Software**

Our outbreak simulation code made use of the SimPy (Vignaux et al., 2012) python module. The RngStreams C library (L’Ecuyer et al., 2002) allowed for the simulations to run in parallel. We used the AD Model Builder (Fournier et al., 2011) and R2admb (Bolker and Skaug, 2011), an R (Development Core Team, 2010) interface for it, to optimize the log likelihood. We prepared graphics with the R package ggplot2 (Wickham, 2009). Code capable of reproducing the results is available from the authors on request.

**Results**

We first present a simulation study that illustrates the accuracy of our methods on a large number of small outbreaks that were simulated from our model. With that self-consistency test passed, we then present estimates from the application of our method to the norovirus data. To show the extent to which our model was appropriate for the norovirus data, we then present diagnostics of the fit.

**Simulation**

We used simulation to see how many outbreak data sets may be required for estimates to be approximately normally distributed around the true parameter values. The simulations also allowed us to gauge the effects of the imputation and rounding necessary for our application.

Much previous work has shown that estimation with data from a single, small outbreak is unreliable (Shao, 1999, and refs. therein). Thus one benefit of aggregating data from multiple outbreaks is that it allows for data from minor outbreaks to produce reliable estimates. However, using data from minor outbreaks does represent a worst-case scenario in the sense that each such outbreak contributes only a small amount of information. For those two reasons, and to keep the simulation study at a manageable size, we restricted our simulations to one set of parameters that is guaranteed to result in small outbreaks. To allow for comparison with our fits to the norovirus data, we used the parameters estimated for our baseline regression group.

As expected, the estimates were not very good when using data from single outbreaks (Fig. 1). In about 49% of these simulations, the initial infective failed to infect anyone, limiting estimation to the length of the symptomatic period. In about 13% of these simulations, only one transmission occurred and the transmission and growth rate parameters were unidentifiable. In about 21% of these outbreaks, the estimate of \( r \) was on the lower bound of parameter space, preventing calculation of Wald confidence intervals. In the remaining 17% of replicates, the coverage probability of the
Table 2
Simulation results for transmission rate $\beta$ and initial growth rate $r$. $n$ denotes the number of outbreaks simulated for an estimate. Imputed refers to the fraction of recovery times deleted and then imputed as described in the Methods section. Rounding refers to the fraction of case records deleted before fitting the data. Rounded indicates whether the onset and recovery times were rounded to whole days. In the simulations, $\beta$ was set to 0.0037 transmissions per infective-susceptible day and the $r$ was set to 0.2664 transmissions per infective day.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Imputed</th>
<th>Missing</th>
<th>Rounded</th>
<th>Bias($\beta$)</th>
<th>Av. s.e.($\hat{\beta}$)</th>
<th>$\beta$ cover. (%)</th>
<th>Bias($r$)</th>
<th>Av. s.e.($r$)</th>
<th>$r$ cover. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0</td>
<td>0.198</td>
<td>0.131</td>
<td>82</td>
<td>0.52</td>
<td>0.834</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0</td>
<td>0.21</td>
<td>0.129</td>
<td>0.81</td>
<td>81</td>
<td>0.55</td>
<td>0.84</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>0.53</td>
<td>0.00</td>
<td>0.251</td>
<td>0.132</td>
<td>0.8</td>
<td>80</td>
<td>0.65</td>
<td>0.966</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.05</td>
<td>0.234</td>
<td>0.137</td>
<td>0.8</td>
<td>81</td>
<td>0.63</td>
<td>0.98</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.00</td>
<td>0</td>
<td>0.214</td>
<td>0.134</td>
<td>81</td>
<td>0.61</td>
<td>0.96</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.05</td>
<td>0.231</td>
<td>0.149</td>
<td>0.8</td>
<td>83</td>
<td>0.66</td>
<td>0.96</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.53</td>
<td>0.00</td>
<td>0.234</td>
<td>0.137</td>
<td>80</td>
<td>0.65</td>
<td>0.966</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.214</td>
<td>0.134</td>
<td>0.81</td>
<td>81</td>
<td>0.61</td>
<td>0.96</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.00</td>
<td>0</td>
<td>0.0195</td>
<td>0.0456</td>
<td>89</td>
<td>0.0389</td>
<td>0.2882</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0</td>
<td>0.017</td>
<td>0.0457</td>
<td>90</td>
<td>0.0378</td>
<td>0.2888</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.53</td>
<td>0.00</td>
<td>0</td>
<td>0.0168</td>
<td>0.055</td>
<td>92</td>
<td>0.0538</td>
<td>0.319</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.017</td>
<td>0.055</td>
<td>0.0562</td>
<td>93</td>
<td>0.0587</td>
<td>0.319</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.00</td>
<td>0</td>
<td>0.00055</td>
<td>0.0067</td>
<td>94</td>
<td>0.0027</td>
<td>0.0738</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0</td>
<td>0.00053</td>
<td>0.00614</td>
<td>94</td>
<td>0.0025</td>
<td>0.0737</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.53</td>
<td>0.00</td>
<td>0</td>
<td>0.00066</td>
<td>0.00617</td>
<td>93</td>
<td>-0.0004</td>
<td>0.0753</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.00067</td>
<td>0</td>
<td>0.00668</td>
<td>96</td>
<td>0.0195</td>
<td>0.08036</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0</td>
<td>-0.00018</td>
<td>0.00724</td>
<td>97</td>
<td>0.0174</td>
<td>0.08222</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>-0.00024</td>
<td>0.00707</td>
<td>97</td>
<td>0.0103</td>
<td>0.08018</td>
<td>87</td>
<td></td>
</tr>
</tbody>
</table>

95% Wald confidence intervals ranged from 80 to 90% (Table 2) and the bias and average standard error for the transmission rate was almost 100 times the true value of the parameter. The average correlation between the transmission rate and initial growth rate estimates was 94%. Estimates for the symptomatic period, although obtained for all replicates, were also not accurate (Fig. 1 and Table 3).

Rounding, deleting 5% of case records, and imputing 53% of recovery times all generally increased the average standard error of estimates, with effects in that order. Effects on the bias were somewhat more variable, but the asymptotic effects of these procedures on the bias appears to be zero. However, even in the 100-outbreak scenario the imputation caused coverage probabilities for $r$ to deviate by as many as 13 percentage points from 95% (Table 2), which recommends the use of confidence intervals that account for the imputation, such as the ones we used in our application.

On the whole, the estimates were much more accurate in the 10- and 100-outbreak scenarios (Fig. 1, Tables 2 and 3). They were also more robust. Estimates for $r$ were on the lower bound 5% of the time in the 10-outbreak scenario and never on the lower bound in the 100-outbreak scenario. The likelihood was divergent about 7–10% of the time in the 10-outbreak scenario versus 0.1–2% of the time in the 100-outbreak scenario. The average correlation between the estimated transmission rate and growth rate was about 0.83 and 0.74 for replicates in the 10- and 100-outbreak scenarios, respectively.

In sum, the method works well with a sufficiently large data set. Moderate amounts of imputation, missing data, and rounding will have mostly modest effects on estimates. Simulation, as part of a parametric bootstrap procedure, can provide an indication of the accuracy of estimates for a particular data set of interest. We demonstrate such a procedure in our application.

Estimates for norovirus in health-care settings

We fitted our generalized linear model to data from a large prospective study of gastroenteritis in health-care settings (Lopman et al., 2004b). In this one-year study, patients and the care staff assigned to any of about 4500 beds in health-care facilities in the former County of Avon, England, were under active surveillance. Trained staff members recorded the dates over which people

Table 3
Simulation results for symptomatic mean $\mu$ and dispersion parameter $\rho$. $n$ denotes the number of outbreaks simulated for an estimate. Missing refers to the fraction of case records deleted before fitting the data. Rounded indicates whether the onset and recovery times were rounded to whole days. Cover refers to the coverage probability of Wald confidence intervals. Lower $\hat{\beta}$ and upper $\hat{\beta}$ refer to the bounds of a bootstrap confidence interval. In the simulations, $\mu$ was set to 3.32 days and the $\rho$ was set to 0.58.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Missing</th>
<th>Rounded</th>
<th>Bias($\hat{\mu}$)</th>
<th>Av. s.e.($\hat{\mu}$)</th>
<th>Cover. (%)</th>
<th>Bias($\hat{\rho}$)</th>
<th>Lower $\hat{\rho}$</th>
<th>Upper $\hat{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>0</td>
<td>-0.60</td>
<td>3.76</td>
<td>83</td>
<td>-0.067</td>
<td>0.01</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0</td>
<td>-0.65</td>
<td>4.13</td>
<td>80</td>
<td>-0.096</td>
<td>0.01</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>0.53</td>
<td>0.00</td>
<td>-0.104</td>
<td>4.26</td>
<td>80</td>
<td>-0.027</td>
<td>0.00</td>
<td>2.00</td>
</tr>
<tr>
<td>10</td>
<td>0.00</td>
<td>0</td>
<td>-0.104</td>
<td>1.51</td>
<td>91</td>
<td>-0.005</td>
<td>0.31</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0</td>
<td>-0.106</td>
<td>1.34</td>
<td>91</td>
<td>0.008</td>
<td>0.32</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>0.53</td>
<td>0.00</td>
<td>-0.106</td>
<td>2.12</td>
<td>90</td>
<td>-0.016</td>
<td>0.23</td>
<td>1.12</td>
</tr>
<tr>
<td>100</td>
<td>0.00</td>
<td>0</td>
<td>-0.010</td>
<td>0.4703</td>
<td>94</td>
<td>-0.0006</td>
<td>0.49</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0</td>
<td>-0.007</td>
<td>0.853</td>
<td>94</td>
<td>-0.0012</td>
<td>0.46</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>0.53</td>
<td>0.00</td>
<td>-0.010</td>
<td>0.6928</td>
<td>94</td>
<td>0.0132</td>
<td>0.47</td>
<td>0.76</td>
</tr>
</tbody>
</table>
were symptomatic and took samples that allowed for laboratory confirmation of the causes of outbreaks. Fig. 2 shows the case histories that were used to fit our model.

The predictors in our model were facility type, which indicated whether an outbreak took place in a long-term-care facility (LTCF) or a hospital; size class, which indicated the number of patients and staff in the unit; period, which indicated the time of the study year when the outbreak began; and case type, which indicated whether a case was a patient or a member of the care staff. See the Methods section for more details.

For our baseline regression group of patients in large hospitals in the fall and winter, the estimate (95% bootstrap confidence interval) of the transmission rate was 0.0037 (0.0026–0.0052) transmissions per infective-susceptible day, that of the initial growth rate was 0.27 (0.23–0.30) transmissions per infective day, that of the symptomatic period was 3.35 (3.09–3.57) days, and that of the dispersion parameter $\rho$ for the symptomatic period was 0.57 (0.54–0.65). Those parameter estimates have been transformed from the log scale for ease of interpretation. The full set of untransformed estimates is given in Table 4.

The basic reproduction number $R_0$ is the expected number of new infections that a single infection will cause at the beginning of an outbreak. Table 5 contains values of $R_0$ calculated from the regression coefficients using the formula $R_0 = (r_{patient} \cdot r_{staff} \cdot \mu_{patient})$.

![Fig. 1. Estimates versus number of outbreaks. The row names indicate parameters. Each small black point represents an estimate. The larger gray points represent the means of the estimates. The horizontal lines represent the values of the parameters used to simulate the data.](image1)

![Fig. 2. Case histories. Each horizontal bar represents the history of a person. The symptomatic period is filled in. Case IDs were assigned by sorting the cases first by onset time, then by recovery time, and then by a random ordering. Initial infectives were given negative case IDs. The panels are arranged so that the outbreak size increases from top to bottom and the outbreak length increases from left to right. Case histories from long-term-care facilities (LTCFs) are in light gray. Some of the times of recovery from symptoms were imputed as described in the Methods section.](image2)

![Fig. 3. Shows estimates for the effect on model parameters of an outbreak being different from the outbreaks in the reference group with respect to one of our predictive variables. The largest effects are the increase in transmission and growth rates in long-term-care facilities (LTCFs) and the reduction in these rates in staff. It appears that transmission rates are higher in the smaller units. Symptomatic periods were estimated to be about 25% shorter for outbreaks in LTCFs and 20% shorter for cases among staff.](image3)

### Diagnostics

As a general test of model fit for the transmission rate and growth rate likelihoods, we calculated the percentile of the log likelihood of the fit to the real data in the distribution of log likelihoods generated by bootstrapping. Out of 10,000 bootstrap replicates, our optimization code found estimates in 9809 cases. The log likelihood of the fit to the real data was in the 25th percentile of the log likelihoods from these estimates. Thus, the log-likelihood of our fit to the real data is not extreme, consistent with a good model fit.

Our use of the moments estimator for the dispersion parameter $\rho$ in (1) precluded a similar assessment of model fit for the

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Regression coefficients (95% bootstrap confidence interval) for the natural log of the outbreak-model parameters.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transmission rate</td>
</tr>
<tr>
<td>Intercept</td>
<td>-5.60 (-5.96, -5.25)</td>
</tr>
<tr>
<td>LTCF</td>
<td>1.98 (1.61, 2.35)</td>
</tr>
<tr>
<td>Small</td>
<td>0.45 (0.34, 0.70)</td>
</tr>
<tr>
<td>Spring-summer</td>
<td>-0.06 (-0.29, 0.14)</td>
</tr>
<tr>
<td>Staff</td>
<td>-1.03 (-1.43, -0.74)</td>
</tr>
<tr>
<td>Dispersion</td>
<td></td>
</tr>
<tr>
<td>Num. obs.</td>
<td>1523</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-1625</td>
</tr>
</tbody>
</table>
symptomatic periods. However, inspection of the default diagnostic plots for glm objects in R did not indicate any problems.

The estimated value of the log of the initial numbers of susceptibles (calculated as \( \log \hat{X}_i^{(0)} = \log(\hat{r}_i / \hat{h}_i) \)) provided a means for a sanity check of our estimates. As described in the Methods section, our size-class predictive variable was determined from the number of beds or staff assigned to a unit. We chose not to directly use the numbers of beds or staff as the initial number of susceptibles in our model because they are likely noisy measurements of the true value. However, it is worth noticing in Fig. 4 that our estimates are on approximately the same scale as the numbers of beds and staff. Furthermore, our estimates replicate two qualitative differences seen in the bed-number and staff-size data. First, units in the small size class do indeed have smaller \( \log \hat{X}_i^{(0)} \) than do units in the large size class. Second, \( \log \hat{X}_i^{(0)} \) for staff are larger than those for patients. On the other hand, \( \log \hat{X}_i^{(0)} \) is larger for hospitals than for LTCFs while LTCFs have more beds and larger staffs. This difference might be reduced by increasing the complexity of our model, but such an exercise in model selection is beyond the scope of this paper.

Table 5: Predicted \( R_0 \) values (95% bootstrap confidence interval) based on regression coefficients.

<table>
<thead>
<tr>
<th>Facility</th>
<th>Season</th>
<th>Size class</th>
<th>( R_0 )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTCF</td>
<td>Fall–winter</td>
<td>Small</td>
<td>2.78</td>
<td>(2.55, 3.23)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>2.42</td>
<td>(2.00, 2.98)</td>
</tr>
<tr>
<td></td>
<td>Spring–summer</td>
<td>Small</td>
<td>2.90</td>
<td>(2.69, 3.29)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>2.74</td>
<td>(2.53, 3.04)</td>
</tr>
<tr>
<td>Hospital</td>
<td>Fall–winter</td>
<td>Small</td>
<td>1.27</td>
<td>(1.22, 1.43)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>1.20</td>
<td>(1.14, 1.33)</td>
</tr>
<tr>
<td></td>
<td>Spring–summer</td>
<td>Small</td>
<td>1.33</td>
<td>(1.26, 1.53)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Large</td>
<td>1.25</td>
<td>(1.19, 1.41)</td>
</tr>
</tbody>
</table>

To graphically evaluate the fit of our model, we plotted observed values of cumulative exposure before infection, symptomatic period duration, and cumulative incidence against those expected from the fitted model (Figs. 5, 6, and 7).

Fig. 3. Regression effect estimates. The column names indicate levels of the categorical predictive variables by which outbreaks differed from the reference group. The reference group was made up of all cases of norovirus among patients in large hospitals in the fall and winter. The row names indicate parameters. The histograms display the distributions of estimates obtained in a parametric bootstrap. Gray rectangles indicate a 95% confidence interval based on the percentiles of bootstrap estimates. The white horizontal line inside each rectangle indicates the ML estimate. LTCF stands for long-term-care facility.

Fig. 4. Estimated initial susceptibles and measures of care-unit size. The column names, row names, and x-axis marks indicate the levels of the predictive variables that determined the estimate of initial susceptibles represented by the gray bars. The middle 95% of estimates from the parametric bootstrap were used to determine the upper and lower bounds of the gray bars. Each point corresponds to an observed outbreak of NoV and indicates either the number of beds for patients in the care unit or the number of staff assigned to a unit. LTCF stands for long-term-care facility. Estimates of initial susceptibles did not vary much by the period predictive variable, so estimates and data from both periods are grouped together in this plot.

Fig. 5. Quantile–quantile plots of the cumulative exposure of people at the time of infection. The semi-transparent lines connect data points from the same outbreak. The opaque diagonal line is a reference line that indicates where points from perfectly matching distributions would fall.
is visible as the high density of points below the reference line. For high values of cumulative exposure, the observed distributions seem to be truncated at around 220 infective-person days, whereas the fitted model assumes an exponential distribution that predicts much larger values.

In Fig. 6, observed symptomatic periods are generally close to predicted symptomatic periods. However, there seems to be a tendency for symptomatic periods in some outbreaks to be less varied than the model predicts. This tendency is visible as lines that start above the reference line on the left side of the plot and then bend to the right quickly enough to be under the reference line by the time they end.

In Fig. 7, cumulative incidence is higher than expected early in outbreaks as a consequence of the larger-than-predicted number of cases with small exposure (Fig. 5). The difference between observed and expected cumulative incidence tends to become less positive or slightly negative by the end of outbreaks, and the absolute difference is usually less than 10 throughout.

**Discussion**

We have shown that reliable estimation of parameters from many small outbreaks is possible using a generalized linear model based on standard stochastic epidemic models. A simulation study demonstrated that we are able to accurately estimate parameters when the data stem from small outbreaks even when some data are missing and about half of recovery times are imputed. Fitting the model to a large number of outbreaks of norovirus, we found that facility type, facility size, and case type seem to have significant effects on outbreak dynamics.

Höhle (2009) described a highly general formulation of stochastic epidemic models within a regression framework, and our model is almost a special case of that general formulation. However, that formulation did not include regressions for the mean infectious period or the initial growth rate, which we include here. Modeling details aside, our work here differs from Höhle (2009) and related regression approaches (Forrester and Pettitt, 2005; Voirin et al., 2011; Meyer et al., 2012) in demonstrating the particular value of a multiple-outbreak regression when fitting data from small outbreaks, which are quite common in health-care settings.

The most striking result of our regression estimates (Fig. 3) are the approximately 7-fold increase in transmission rates and 3-fold increase in initial growth rates in the long-term-care facilities (LTCFs) relative to hospitals. Fig. 2 shows that LTCF outbreaks do indeed include many of the larger and faster growing outbreaks in the data set.

The higher transmission rates for occupants of LTCFs may be a consequence of occupants having more opportunity to socialize in large groups. Alternatively, we may be seeing the effects of our assumptions of a closed population and homogeneous mixing being violated. Hospitals have more rapid turnover of patients, and the exposure of people who arrived in the care unit after the outbreak started will be overestimated in our model. Occupants of LTCFs may vary more in contact rates by virtue of behavioral differences, and such variation in exposure could lead to a higher initial growth rate (Becker, 1989, pp. 133–138).

Our model is agnostic about the particular pathways of transmission in the outbreaks, and thus the estimates represent rates of transmission by all routes including transmission by person-to-person, environmentally-mediated, and foodborne routes. Outbreak investigations often associate foodborne transmission with an abrupt increase in infections (e.g., Isakbaeva et al., 2005), and thus difference in the frequency of this mode of transmission between settings could cause differences in our estimates. That being said, we did not include infection terms corresponding...
to foodborne transmission because most reported norovirus outbreaks are attributed primarily to person-to-person transmission (Yen et al., 2011), and we believed that all the outbreaks we analyzed involved primarily a combination of person-to-person and environmentally-mediated transmission. In crowded settings such as healthcare-facilities, it is often difficult to determine the relative frequency of these routes of transmission (Kuusi et al. 2002; Isakbaeva et al., 2005; Centers for Disease Control and Prevention, 2008).

Although our aim was not to fit the data to a highly realistic model, we did take a significant step towards realism by modeling variation in the initial growth rate of the outbreak. The transmission rate in our model determines how the expected number of new cases increases as the product of the numbers of infectives and susceptibles increases. Thus estimates of the transmission rate will be highly sensitive to those of the initial number of susceptibles, which determines the number of susceptibles throughout the outbreak. But in the case of norovirus, the number of susceptibles is difficult to know as there is no serological correlate of protection. Recent work on joint estimation of transmission rates and the initial number of susceptibles with data from a single outbreak (Hayakawa et al., 2003; Huggins et al., 2004; Lau and Yip, 2008; Kypraisios, 2009) has shown that estimates of the initial number of susceptibles tend to be low when data sets are small. As we have seen in our simulation results (Fig. 1), this bias decreases as the number of outbreaks in the data set increases, even if all outbreaks are small.

The estimates for NoV transmission dynamics we calculated complement results from previous epidemiological analyses of NoV in health-care settings. Previous analyses of our data set (Lopman et al., 2004a,b) had examined how risk of NoV infection or particular symptoms of NoV infection varied with age and other personal characteristics. The current analysis adds to these results by quantifying effects that could be used to predict norovirus outbreak dynamics.

Analysis of a 2003–2006 study of NoV outbreaks in long-term care facilities (LTCFs) in Oregon (Rosenthal et al., 2011) suggested that larger facilities may have a higher risk of experiencing outbreaks. Our result that transmission rates are lower in larger facilities suggests that any increased risk that larger facilities have is not caused by increased transmission rates. However, our result must be interpreted with caution because we have not been able to account for many factors that may affect contact rates, such as number of beds per room. Whether the cause of the lower transmission rates is really facility size or a correlation of facility size with some omitted variable is unclear.

A few previous studies have estimated individual-level parameters for NoV that are comparable to our estimates. Using data from a NoV outbreak in a primary school and nursery in Derbyshire, England, O'Neill and Marks (2005) estimated that the probability of a susceptible person avoiding infection from an infective person in the school for a day was 0.998. Using the formula \( \Pr(\text{avoidance}) = \exp(-\beta \times 1 \text{ susceptible} \times 1 \text{ infective} \times 1 \text{ day}) \), our estimates yield \( \Pr(\text{avoidance}) \) that ranges from about 0.995 for patients in small LTCFs to 0.999 for staff in large hospitals.

Hejne et al. (2009) estimated the basic reproduction number of NoV in boy-scout camps to be about 14 and 7, respectively, under two different sets of assumptions. Our highest \( R_0 \) was approximately equal to 3 (Table 5). The relative lowness of our \( R_0 \)s might reflect contact rates being higher in the camp setting, and it may also reflect the effect of better hygiene in the health-care settings. Hejne et al. (2009) estimated that the implementation of an enhanced hygiene protocol drove the reproduction number in the camps down to about 2 and 1, values on par with our own estimates. Likewise, Hejne et al. (2012) estimated reproduction numbers for an outbreak within the wards of a psychiatric institution to be close to one and thus close to our hospital \( R_0 \).

Actually, those estimates tend to be lower than ours, but they are reproduction numbers averaged over the full course of an outbreak and thus should be expected to be lower due to the depletion of susceptibles over time. Our estimates may be more generalizable than the both the psychiatric-ward and boy-scout estimates because our data set was larger and included data from both large and small outbreaks.

Zelner et al. (2010) used data from a Stockholm outbreak to estimate that the average infectious period was 1.2 days. The setting of these outbreaks was households that included children in daycare centers. Thus, the infectious period may have been shorter in these outbreaks because many of the infectives were likely healthy people between the ages of 5 and 70, whereas people below the age of 5 and, to an even greater degree, people over the age of 70 were over-represented in our data (Lopman et al., 2004b). In our data, people in these extreme age groups had average symptomatic periods of 3 days (Lopman et al., 2004b). Moreover, the Stockholm estimate is based on reported infectious periods rather than symptomatic periods, which were not reported. As a result, if the assumed initial number of susceptibles for the Stockholm analysis was too high, the infectious period would have been underestimated.

Although our estimates of the symptomatic period may be relatively long, it is possible that some of the patients were discharged into the community before they became asymptomatic. Thus, for patients, our estimates most accurately describe the period of being symptomatic while simultaneously being in a health-care facility.

The daily transmission rates estimated from the Stockholm data, 0.14 transmissions per infective-susceptible day, are more than 3-fold higher than our highest estimated transmission rate, which was 0.04 transmissions per infective-susceptible day for patients in small LTCFs. The joint estimation approach we used could be applied to the Stockholm data to determine whether the higher transmission-rate estimates may have resulted from underestimation of household sizes.

However, the transmission rates may well be different because of differences in hygiene measures, contact rates, or different levels of baseline immunity. Nurses for example may be frequently exposed and therefore highly immune. Additionally, time-series analysis of outbreak occurrence (Lopman et al., 2009) has suggested that transmission rates generally may vary with host, weather, and virus factors. Taken together, these differences may explain the large discrepancy in estimated transmission rates. The compilation and analysis of a large, multiple-outbreak data set that includes predictive variables indicative of hygiene, contact rates, and baseline immunity could shed light on which of these elements has the greatest effect on transmission rates.

In our application, we made the simplifying assumption that the latent period was fixed at its mean, which allowed us to directly calculate infection times from the reported onset of symptoms. The infection time determines the cumulative exposures \( h_j \) in (3). Because the cumulative exposure is a non-linear function of time and the mean of a non-linear function of a random variable does not always equal the function evaluated at the random variable\'s mean, the extent to which latent periods varied in reality likely introduced bias into our calculated cumulative exposures and the estimates based on them. The bias could be either positive or negative depending on whether cumulative exposure usually increases more quickly before or after the assumed transmission times.

Another simplifying assumption we made was that people were only infectious when they were symptomatic. In support of this assumption, Sukhrie et al. (2012) have shown that asymptomatic people are much less infectious than symptomatic people in health-care settings. To quantify the effect of this and the fixed latent period assumption, we could make the infectious period a latent variable that we integrate over to evaluate the
likelihood, as in Hohle et al. (2005). Alternatively, we could use a kernel-smoothing method to estimate unobserved latent or infectious periods, as in Lau and Yip (2008). However, even without such calculations it is clear that if, in reality, the infectious period extends beyond the symptomatic period, our estimates of transmission rates have been inflated by our underestimation of exposure.

From the numerical results displayed in Tables 2 and 3, we see that highly reliable estimation depends on collection of an extensive data set. The Centers for Disease Control and Prevention (CDC) has recently established a National Outbreak Reporting System that, with the contributions of state health departments, will provide more comprehensive surveillance for all U.S. gastroenteritis outbreaks (Centers for Disease Control and Prevention, 2011). However, the data we have analyzed here is more detailed than what is routinely collected in outbreak investigations. More detailed outbreak investigations are needed to collect such data and further characterize modes of transmission. The collection of NoV genomic data may also be of great value (Teunis et al., 2013).

Regarding the general subject of outbreaks of hospital-acquired infections in individual hospitals, we suspect that in some hospitals a large part of the necessary data collection is already taking place as a part of existing surveillance programs. A 2008–2009 survey estimated that approximately one third of California hospitals used automated surveillance technology to monitor hospital-acquired infections (Halpin et al., 2011). Similarly, a survey of hospitals in the Northeastern United States found that one third of hospitals had an electronic surveillance system in place (Grota et al., 2010). Such systems were used to detect outbreaks, analyze data, and generate reports of hospital-acquired infections (Grota et al., 2010). As several states mandate reporting infection rates of MRSA and many more require some form of reporting of hospital-acquired infections (Committee to Reduce Infection Deaths, 2011), many hospitals may have data on the total number of cases in many outbreaks for several pathogens of concern.

Our analysis, based on a robust data set, demonstrates that parameter estimates are substantially less biased when a large number of outbreaks are fitted. We submit that, for norovirus and many other pathogens, there are several uses for accurate estimates of transmission rates, initial growth rates, and infectious periods. Policy-makers can use such estimates to compare the efficacy of different control strategies such as hygiene protocols, isolation measures, prophylactic treatments, and vaccination policies. Those monitoring the small outbreaks of zoonotic diseases may be able to use such estimates to identify variables that make transmission more likely.

Role of the funding source

This work was supported by NSF Grant EF-0742373 and NIH Grant R01 GM088344. E.B.O. was also supported by BEACON Grant 61-2075UT. K.M.P. was also supported by the RAPIDD program of the Science and Technology Directorate, U.S. Department of Homeland Security, and the Fogarty International Center, NIH. Our funding sources had no role in this study’s design; in the collection, analysis, and interpretation of the analyzed data; in the writing of this paper; or in the decision to submit this paper for publication.

Acknowledgments

The Texas Advanced Computing Center at UT provided computing resources.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.epidem.2013.12.002.

References


Centers for Disease Control and Prevention, 2011. Updated norovirus outbreak management and disease prevention guidelines. MMWR 60, 1–18.


gastroenteritis in a rehabilitation centre due to environmental contamination. Epidemiology and Infection 129, 133–138.


