The Milwaukee Cryptosporidium outbreak: assessment of incubation time and daily attack rate
Mukul Gupta and Charles N. Haas

ABSTRACT
The time course of reported illnesses (epidemic curve) in the 1993 Milwaukee outbreak of cryptosporidiosis was analysed using a dynamic model considering time variant force of infection and incubation time distributions. Different functional forms for the force of infection and incubation time distribution were tested. The resulting model is a coupled integro-differential equation system. These models gave a good fit to the data, although depending upon the functional forms of the underlying distributions, different incubation time and force of infection curves were obtained. However there was reasonable agreement with respect to a baseline illness rate that existed. This demonstrates that useful information may be obtained in this manner, although it should be supplemented with other data (e.g. serology) for a precise assessment of dynamics of disease occurrence during waterborne epidemic conditions.

Key words | Cryptosporidium, drinking water, mathematical epidemiology, Milwaukee

INTRODUCTION
In the spring of 1993 there was a widespread outbreak of cryptosporidiosis among the residents of Milwaukee, Wisconsin. A 100-fold increase in the rate of isolation of Cryptosporidium was noted and an estimated 403,000 people became ill (Mac Kenzie et al. 1994). A telephone survey was conducted to ascertain the date of onset of symptoms. However the precise duration of exposure, as well as the estimated magnitude of exposure, has only been crudely estimated. The objective of this work is to present a mathematical model for that outbreak and to demonstrate whether and how the attack rate, exposure and incubation time distribution can be estimated from the epidemic curve.

BACKGROUND
Prior work on incubation times
The incubation time, as defined here, is the interval between an individual acquiring the agent of disease and the onset of symptoms. In this sense, the definition is of the duration of the latent period, as defined in Anderson & May (1991). In a given population exposed to an agent, there will be a distribution of incubation times. While there has been a rich literature on incubation times, it has been restricted primarily to incubation after person to person exposure (Bailey 1975; Becker 1989). In such a case (typified by household contacts), the precise times between the onset of contiguity in the index case and the onset of symptoms in subsequent cases is more readily ascertained. Only in a few cases, such as in transfusion-acquired AIDS, has the epidemic curve (data on cases versus time) been used to back-calculate likely exposure and incubation times (Medley et al. 1988).

Using primarily household data, and also information from defined discrete exposures, Sartwell (1950) tabulated the means and standard deviations of incubation times for a diversity of infections using the lognormal distribution. However, his method of ‘fitting’ was not documented.
In work done in which random walk models of the process of microbial birth and death were depicted within a host, Williams (1965a,b) developed an analytical form for the distribution of incubation times in terms of the initial challenge, and the in vivo birth and death rates for the pathogen. This model had interesting consequences, including a specific prediction of an inverse relationship between mean incubation time and inoculum dose at low doses. This distribution has the right skewed form similar to a lognormal. Although subsequent experimental studies in plants (Ercolani 1985) and animals (Seto & Takizawa 1969) have at least qualitatively supported these results, the work of Williams has not apparently been substantially elaborated upon.

When population models have been used to study environmentally acquired infection (here using the term ‘environmental’ to encompass any source of infection other than person to person), incubation time distributions have been modelled either as exponential or gamma distributions, for example, Eisenberg et al. (1996). This has been done largely for computational simplicity since a gamma distribution can be modelled as a series of latent stages in sequence. This approach is largely the one taken in modelling dynamics in the HIV/AIDS situation, for example, see Cairns (1995).

**METHODS**

**Epidemic state model**

The compartment modelling approach has been used extensively for modelling disease transmission through human populations. The target population can be segmented into groups, and the rate of individuals moving between groups described. The form of the model is a set of differential equations describing system dynamics. The genesis of this formulation stems from Kermack & McKendrick (1927, 1932, 1933). This approach has been documented in a number of monographs (Bailey 1975; Becker 1989; Anderson & May 1991).

In the problem at hand, there are four states that are of potential interest. These are the susceptible (or naïve) state of uninfected individuals who are capable of being infected; the latent state consisting of individuals who have been exposed to *Cryptosporidium* and who will ultimately become infected or ill; the symptomatic state and the asymptomatic state (consisting of individuals who are infected, but who do not evidence signs of illness). The transitions between these compartments are shown diagrammatically in Figure 1.

The differential equations, for the compartment model summarized in Figure 1, are:

\[
\frac{dx(t)}{dt} = -\beta(t)x(t) \tag{1}
\]

\[
\frac{dy(t)}{dt} = \beta(t)x(t) - Q(t) - R(t) \tag{2}
\]

\[
\frac{dz(t)}{dt} = Q(t) \tag{3}
\]

\[
\frac{dI(t)}{dt} = R(t) \tag{4}
\]

In general only the number of newly symptomatic individuals are observed and epidemiological records show the number of new notifications each day or week. We therefore use the concept of an epidemic curve giving the rate, \(w\), at which new cases are recognized, i.e.
\[ w = \frac{dz(t)}{dt} = Q(t) \]  

(5)

**Simplified compartment model**

If information is only desired on the number of symptomatic cases, or the rate at which new symptomatic cases develop, the above model may be simplified into two differential equations:

\[ \frac{dX(t)}{dt} = -\beta(t)X(t) \]  

(6)

\[ \frac{dZ(t)}{dt} = Q(t) \]  

(7)

This simplification is possible since the number of asymptomatic cases \( Y \) does not appear in either Equations (6) or (7), or (as will be discussed below) the functions \( Q(t) \) and \( \beta(t) \).

**Rates**

The transition of individuals from susceptible to latent and thence to ill is given by two rate functions, \( \beta(t) \), and \( Q(t) \). These are, respectively, the force of infectivity (Anderson & May 1991) in units of 1/time, and the rate of new symptomatic cases (in individuals per day). The formulation of these rates is described below.

**Force of infectivity**

At a given time \( t \), the force of infectivity is given as \( \beta(t) \), so that \( \beta(t) \times (t) \) is the instantaneous number of new infections per unit time added to the infected group at time \( t \). In the case of a point outbreak with a single exposure of short duration, \( \beta(t) \) would be an impulse function (Dirac delta) whose integral indicates the overall proportion of individuals who ultimately will become, either symptomatically or asymptotically, infected. This may be generalized to include the potential for a baseline by the following equation:

\[ \beta(t) = b_0 + b_1g(t) \]  

(8)

where: \( \beta \) is the force of infection (1/time); \( g(t) \) is the proportion of susceptibles ultimately becoming infected who will become infected in the interval \( <t,t+dt> \); \( b_0 \) is the background infectivity rate; and \( b_1 \) is the total proportion of susceptibles who will become infected during the incident (strictly speaking, this is a theoretical number for an infinite population, since as will be noted in the fitting, the buildup of infected persons is so great that for some choices of distributions \( f(t) \) and \( g(t) \), the latter stages of the outbreak are diminished not so much by a diminution in the force of infection, but a diminution in the size of the remaining susceptible population).

Defined in this manner, \( g(t) \) may be any probability density function defined over the continuous positive real line. This density function may contain one or more parameters defining the location, scale and shape of the time dependency of the force of infection. This may also relate to the distribution (with time) of the dose of pathogens to which a population may be exposed. The term \( b_0 \) accounts for any baseline or endemic rate of illness, which may not necessarily be associated with the particular risk of interest. This baseline may also include potential risks from secondary (person to person) contacts, providing that they remain relatively constant over the time being investigated.

**Rate of new illnesses**

Our approach to developing \( Q(t) \) follows the rationale of Kermack & McKendrick (1927, 1932, 1933) as also typified by Bongaarts (1989) and DeGruttola & Mayer (1987). The rate of newly ill people at any time \( t \) results in the proportion of persons who became infected at any time \( \tau < \tau \) given the incubation time distribution for elapsed time \( t - \tau \) summed over all \( \tau \). Mathematically, if the incubation time distribution density function is given by \( f(\tau) \), then this may be written as:

\[ Q(t) = \int_{0}^{t} \lambda(\tau)X(\tau)f(t-\tau)d\tau \]  

(9)
In Equation (9), the incubation time distribution may be any probability density function over the positive real line, for example the lognormal, Weibull, gamma or inverse gaussian (Table 1). As noted previously, the gamma distribution may also arise from subdividing the latent period into a series of subcompartments. However the formulation as Equation (9) allows for considerably more latitude in choice of incubation time distributions. As in the case of the force of infection, \( f(t) \) may contain unknown parameters (e.g. the mean and variance of the incubation time) that must be estimated from the data. The parameter \( \lambda \) is the fraction of infected persons who become ill (a fraction of the infected persons, \( 1 - \lambda \), experience only asymptomatic infection).

**Formal model**

The formal model used to depict the Milwaukee outbreak consisted of Equations (6) through (9). This model contains a number of parameters that need to be estimated from the data: \( b_0, b_1, \lambda \), and the parameters of the distributions characterizing the force of infection and the incubation time. Furthermore, the functional forms of the two distributions must be specified.

In this work, four distributions were evaluated as potential candidates to describe the Milwaukee data—the inverse gaussian, Weibull, gamma and lognormal density functions (Table 1). These were used separately for the force of infection, and for the incubation time. Hence a total of 16 combinations (each of the four distributions for incubation time combined with each of the four distributions for the force of infection) were evaluated.

**The Milwaukee data**

The data on cryptosporidiosis experienced during the waterborne outbreak in Milwaukee in 1995 were used. A 100-fold increase in the rate of isolation of Cryptosporidium was noted and an estimated 403,000 people became ill (Mac Kenzie et al. 1994) This estimate of attack was based on a telephone survey of 777 persons (of whom 241 met the case definition of 3 days of either diarrhoea—three loose stools in a 24 h period, or vomiting). In detailed studies it was ascertained that the secondary attack rate was low (Mac Kenzie et al. 1995), and hence the focus on primary cases presented here is appropriate. For the purpose of fitting, 12:01 am on 1 March 1993 was defined as \( t = 0 \) (i.e. any cases that had a day of onset on 1 March were regarded as occurring on day 1).

**Model fitting approach**

The epidemic model defined above consists of a pair of equations: an ordinary differential equation (Equation 6) coupled with an integro-differential equation (Equations...
The overall fitting approach involved embedding a numerical solution routine for these equations into an optimisation routine to find the values for the parameters (for particular choices of the distribution for force of infection and incubation times) of the models which minimize an objective function. The objective function minimized is the Poisson deviance (Morgan 1992).

A computer program was developed using MATLAB to optimise model parameters to predict a best-fit epidemic curve. The Statistics Toolbox routines fmins and fminu (using a Nelder–Mead simplex, and a generalized Newton–Raphson approach, respectively) were used for optimisation. The epidemic model was solved at each step of the optimisation using the ode45 routine (which uses a 4th–5th order Runge–Kutta integrator). To evaluate the integral in Equation (9) an iterative approach was employed. For an assumed (by the optimisation engine) set of parameters, it was first assumed that when evaluating Equation (9) that \( X \) (the number of susceptibles) at any time step was equal to its initial value. Integration of the system produced a revised value of the temporal history of \( X \), which was compared to the assumed value. If there was a significant shift in the model solution then the revised \( X \) values were used to re-solve the problem; otherwise convergence was assumed (a \( 10^{-3} \) relative value for the daily number of new cases was used as the test). A flowchart for the procedure is given in Figure 2.

**Figure 2 | Flowchart for evaluation of epidemic model.**

### Initial conditions

The initial conditions (at \( t = 0 \)) to solve the integro-differential equations were taken to be \( x = 777 \) (corresponding to the number of individuals in the telephone survey), \( I = y = 0 \). The assumption of zero asymptomatic and zero ill persons at the onset, along with the implicit assumption of zero immune individuals, was a parsimonious one given the absence of data on these points. However, in view (as noted subsequently) of the apparent observation of an ongoing endemic attack rate, the neglect of initial asymptomatic, ill, and immune individuals might not have been a strictly accurate one. The impact of changing these assumptions would be appropriate for further study.

### RESULTS AND DISCUSSION

#### Effect of alternative distributions

The Milwaukee data were fit to the outbreak model using four distributions for each of the force of infection \( g(t) \) in Equation 8 and incubation time \( f(t - r) \) in Equation 9. The four distributions examined were the lognormal, gamma, Weibull and inverse gaussian. Each of these is defined over the positive real line and contains two parameters and may be completely described by a mean and variance (Table 1). Hence the fitting procedure
involved determining a total of seven parameters: two parameters for each of the force of infection and incubation time distributions, the fraction of infections resulting in illness ($\lambda$), the background force of infection ($b_0$) and the total force of infection ($b_1$).

**Best-fit parameters**

The best-fit parameters are summarized in Table 2 and shown as star plots in Figure 3. A plot of the observed daily attack numbers versus the fitted epidemic curves is shown in Figure 4. There are 28 days of observed attack rates. The critical $\chi^2$ distribution at 21 degrees of freedom (1 = 28 - 7 parameters) is 32.67. Hence, from a goodness-of-fit point of view, all 16 combinations (of infectivity and incubation distributions) result in adequate statistical fit. As shown in Figure 4, there is no substantial
discernable difference in the prediction of the epidemic curve by different combinations of assumed infectivity and incubation functions.

Comparing the values of the parameters among the panel of 16 represented in Table 2, it appears that the different assumed infectivity and incubation distributions result in differences, which in some cases are substantial, in the fitted parameters. This includes the basic parameters of theoretical mean infectivity and incubation time. This phenomenon may be visualized by means of a star plot, as in Figure 3. There is also a correlation between parameters—particularly between the baseline and total infectivities and the symptomatic fraction, and between the mean infectivity time and the mean incubation time (see Table 3).

**Infectivity and incubation time distributions; residual susceptible population**

The impact of the different parameter combinations, which are obtained as the best fits to the panel of 16 distributional combinations, is that there are gross differences between the infectivity and incubation distributions among the fits. In Figures 5 and 6, the force of infection and the incubation time distributions are shown for each of the 16 cases. The curves are identified by letters A through P, such that an increasing letter value represents a larger value of the computed objection function.

Cases C (LN/LN), D (IG/LN) and G (IG/gamma) show notable qualitative differences from the remaining cases. These three fits show a more prolonged force of infection reaching higher peak values (Figure 5). However, these peaks occur sufficiently delayed so that the bulk of the exposed population has already been infected. Furthermore, in these three cases, particularly C and D, the incubation time distribution has a shorter mean, and less dispersion (i.e. is more peaked) as shown in Figure 6.

There is also a substantial difference in the estimated fraction of uninfected individuals at the end stages of the outbreak. This is shown in Figure 7. Cases C, D and G predict virtually complete disappearance of the uninfected fraction by day 50, while the other cases show a ‘stabilization’ (actually not the case, due to the existence of a baseline) at from 3 to 60% uninfected.

However interesting these differences might be, they do not shed much light on which of the 16 members of the panel of distributional combinations is the most plausible, since data on the uninfected fraction are not available. However this does suggest the importance in future large outbreak investigations of attempting to obtain such information (perhaps by random serological testing) to discriminate among competing quantitative hypotheses.

**Consistency with prior knowledge of incubation times**

The panel of 16 distributions depicts a mean incubation time between 2.9 and 12 days. Qualitatively, this is consistent with reported incubation times of 7.2 days by Jokipii & Jokipii (1986) and 5 days by Chappell et al. (1996). Furthermore, in studying individuals who contracted cryptosporidiosis following short, defined, visits to Milwaukee during the outbreak, a mean incubation time of 6.8 days was observed (MacKenzie et al. 1995). On this basis, most of the 16 cases yield plausible estimates. In the case of the LN/LN (C), IG/LN (D), Weibull/gamma (H) and gamma/gamma (K), the mean incubation times of the fitted model appear substantially less than has been previously reported, while in the case of the Weibull/LN (B), IG/Weibull (E), gamma/Weibull (F), LN/Weibull (I), Weibull/Weibull (J) and LN/IG (N) the mean incubation times appear to be substantially greater than has been previously been reported. Hence the most plausible fits would appear to be the Weibull/IG (A), IG/gamma (G), gamma/IG (L), gamma/LN (M), LN/gamma (O) and IG/IG (P) cases. Notably, however, this plausibility screen retains one of the cases (case G) which exhibits a prolonged infectivity distribution (Figure 5).

**DISCUSSION**

These results suggest that the use of an epidemic curve alone does not provide sufficient evidence to differentiate among different combinations of infectivity and incubation time functions. In this regard, there is some resemblance to the well-known issue of using
<table>
<thead>
<tr>
<th>Distributions</th>
<th>Infectivity distribution parameters</th>
<th>Incubation distribution parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (d)</td>
<td>Standard deviation (d)</td>
</tr>
<tr>
<td>Weibull</td>
<td>31.042</td>
<td>3.133</td>
</tr>
<tr>
<td>LN</td>
<td>27.815</td>
<td>2.993</td>
</tr>
<tr>
<td>LN</td>
<td>41.941</td>
<td>9.095</td>
</tr>
<tr>
<td>IG</td>
<td>43.805</td>
<td>1.545</td>
</tr>
<tr>
<td>IG</td>
<td>27.515</td>
<td>0.830</td>
</tr>
<tr>
<td>gamma</td>
<td>25.373</td>
<td>3.545</td>
</tr>
<tr>
<td>IG</td>
<td>41.375</td>
<td>1.417</td>
</tr>
<tr>
<td>IG</td>
<td>31.571</td>
<td>2.904</td>
</tr>
<tr>
<td>LN</td>
<td>26.618</td>
<td>2.695</td>
</tr>
<tr>
<td>Weibull</td>
<td>25.980</td>
<td>2.113</td>
</tr>
<tr>
<td>gamma</td>
<td>32.243</td>
<td>4.138</td>
</tr>
<tr>
<td>gamma</td>
<td>29.983</td>
<td>3.837</td>
</tr>
<tr>
<td>gamma</td>
<td>28.490</td>
<td>3.611</td>
</tr>
<tr>
<td>LN</td>
<td>24.731</td>
<td>3.572</td>
</tr>
<tr>
<td>LN</td>
<td>28.219</td>
<td>3.316</td>
</tr>
<tr>
<td>IG</td>
<td>28.917</td>
<td>0.714</td>
</tr>
</tbody>
</table>

LN—lognormal.
IG—inverse gaussian.
**Table 3** | Spearman correlation coefficients between parameters among 16 distribution combinations

<table>
<thead>
<tr>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$\lambda$</th>
<th>$\mu_{inf}$</th>
<th>$\mu_{inc}$</th>
<th>$s_{inf}$</th>
<th>$s_{inc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_1$</td>
<td>0.9441</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>****</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-0.9206</td>
<td>-0.9735</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>****</td>
<td>****</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{inf}$</td>
<td>0.4618</td>
<td>0.4588</td>
<td>-0.3235</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_{inc}$</td>
<td>-0.2618</td>
<td>-0.3000</td>
<td>0.1471</td>
<td>-0.9500</td>
<td>****</td>
<td></td>
</tr>
<tr>
<td>$s_{inf}$</td>
<td>-0.1676</td>
<td>-0.2441</td>
<td>0.2735</td>
<td>0.0588</td>
<td>-0.1500</td>
<td>****</td>
</tr>
<tr>
<td>$s_{inc}$</td>
<td>0.4324</td>
<td>-0.2441</td>
<td>0.1412</td>
<td>-0.6147</td>
<td>0.4735</td>
<td>-0.2647</td>
</tr>
</tbody>
</table>

**** $<10^{-4}$ $\mu_{inf}$=mean infectivity time.
*** $<0.001$ $\mu_{inc}$=mean incubation time.
** $<0.01$ $s_{inf}$=standard deviation of infectivity time.
* $<0.05$ $s_{inc}$=standard deviation of incubation time.

**Figure 5** | Fitted force of infection distributions (IG: inverse gaussian; LN: lognormal).

**Figure 6** | Fitted incubation time distributions (IG: inverse gaussian; LN: lognormal).

deconvolution to extract information with respect to the distributions being deconvoluted. In prior work, Medley et al. (1988) attempted to assess the infectivity and incubation distributions amongst transfusion-related AIDS cases, and noted that it was not possible to distinguish among different functional forms for the
infectivity and incubation distributions. In their case, however, the problem was somewhat simpler in that it was assumed that the number of non-infected persons remained essentially constant (i.e. the initial number of uninfected individuals is much greater than the total number of ultimate cases), while the present problem requires a formal consideration of the decline in uninfected persons. Hence in the work of Medley et al. (1988), the problem is posed as a straightforward numerical integration, while in the present case the problem is posed as a solution to coupled integro-differential equations (which is numerically more difficult).

Significance of baseline

In all cases, it was necessary to include a baseline parameter \( b_0 \) to get adequate model fitting. This would suggest that in the periods before and after the peak of exposure, a low level, ‘endemic’ exposure to the risk existed. In fact this has been suggested in analysis of emergency room admissions prior to and during the Milwaukee outbreak (Morris et al. 1998).

The panel of 16 models has a range of 428–945/100,000/day for the baseline rate (Table 2). This rate is quite high, for example as compared to the US ‘foodnet’ data (Centers for Disease Control and Prevention 1998). However it should be noted that the underlying epidemiological information used in this analysis did not include only clinically diagnosed cryptosporidiosis, and thus may have reflected cases from other microorganisms (e.g. viruses); although for most other waterborne pathogens the incubation times are less than for Cryptosporidium (with the exception of Giardia). It should also be noted that the underlying epidemiological survey may have detected illnesses resulting from other vehicles than drinking water, such as food or contact with infected persons or animals. Hence the reason for the high apparent baseline risk is unclear. It should also be noted that the baseline risks, if extended to an annual basis would result in an annual incidence of 1.6–3.3 cases per person per year. This is somewhat (although perhaps only by a factor of 2) greater than the total highly credible gastroenteritis rate noted in a population based study by Payment et al. (1991).

CONCLUSIONS

On the basis of this work we conclude that it is possible to apply epidemiological state models to fitting a large outbreak, such as occurred in Milwaukee in 1993. The use of any number of functions for the force of infectivity and the incubation time distribution is clearly possible. However from an epidemic curve alone, it would not appear possible to always obtain uniquely defined infectivity and incubation distributions—and broad areas of model uncertainty (in addition to parametric uncertainty) may remain.

It is clear from this work that ancillary information, such as uninfected proportion, would be useful in reducing the degree of model uncertainty in fitting epidemic curves. While the current model is sufficiently complex such that computation of confidence bounds is not realistically possible, with advances in computer speed the use of bootstrap likelihood-based methods for constructing confidence bounds should be possible.

Finally, this work has confirmed the finding of Morris et al. (1998) that there was a substantial baseline of illness preceding and following the 1993 Milwaukee outbreak of cryptosporidiosis.
ACKNOWLEDGEMENTS

This work was based in part on the PhD thesis of Mukul Gupta at Drexel University. Dr Joan Rose (Michigan State University) and Dr Michael O’Connor (Drexel University) are acknowledged for useful comments on the dissertation. Funding was supplied by the L.D. Betz Professorship in Environmental Engineering.

REFERENCES

Bongaarts, J. 1989 A model of the spread of HIV infections and the demographic impact of AIDS. Statistics in Medicine 8, 103–120.