Research Article

Statistics in Medicine

Received 12 January 2010, Accepted 16 April 2010 Published online in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3968

Introduction and snapshot review: Relating infectious disease transmission models to data

Philip D. O'Neill[*†]

Disease transmission models are becoming increasingly important both to public health policy makers and to scientists across many disciplines. We review some of the key aspects of how and why such models are related to data from infectious disease outbreaks, and identify a number of future challenges in the field. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: epidemic models; infectious diseases; stochastic epidemic models; data augmentation; Bayesian inference; Markov chain Monte Carlo methods

1. Introduction

During the last decade or so, there has been a huge increase in the level of interest in mathematical models for the transmission of infectious diseases. Although such models were once more or less exclusively the preserve of theorists within the mathematical and statistical community, they are now of interest to a much broader range of scientists, and notably also to national and international level policy makers who wish to control and contain existing and future outbreaks.

Although it is impossible to be definitive, two factors seem likely to have contributed to this increased enthusiasm for models of disease transmission. First, the epidemic modelling community has itself become far more engaged with addressing issues of direct interest to public health policy. Examples include evaluation of vaccination strategies (e.g. [1], concerned with the design of HPV vaccination programmes in the USA), quantitative assessment of the effectiveness of proposed control measures (e.g. [2], concerned with the school closures in response to influenza H5N1), evaluation of the effectiveness of control measures as actually employed (e.g. [3], concerned with the control measures for Severe Acute Respiratory Syndrome (SARS)), and estimation of the key epidemiological quantities of a particular epidemic (e.g. [4], concerned with the estimation of characteristics such as the reproduction number for influenza H1N1 in Victoria, Australia). In some countries at least, the outputs from modelling efforts are now a politically acceptable contribution to the decision-making process on relevant policy issues.

The second factor is the advent of substantially greater and accessible computational power. This power has been utilized in two main ways. First, it is now possible to simulate large and complex models, for instance, describing the individual-level behaviour of millions of individuals on a daily basis, which can be used to assess the effectiveness of possible intervention measures (e.g. [5] outlines the use of the so-called agent-based models to understand and predict the spread of pandemics; [6] considers strategies to mitigate against bioterrorist smallpox attacks in Portland, Oregon, USA, via a large-scale simulation). Second, in common with many application areas of statistics, analyses of data sets on infectious disease outbreaks can now be performed using computationally intensive statistical methods such as Markov chain Monte Carlo (MCMC) methods or sequential Bayesian methods (see, e.g. [7–10]).

Our aim in this paper is to provide a short overview of the area of infectious disease modelling in the context of statistical data analysis, and highlight some areas of future challenge. An important caveat is that, since epidemic modelling is itself the subject of an enormous and diverse literature, our review is by no means comprehensive. This caveat applies both to the list of topics discussed and also to the list of references. However, it is intended that the reader who wishes to explore further will find the first steps into the relevant literature.

School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.
[*]*Correspondence to: Philip D. O'Neill, School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.*
[†]*E-mail: philip.oneill@nottingham.ac.uk*

The paper is structured as follows. In Section 2 we begin by reviewing a few of the basic modelling concepts that are common to many epidemic models. Section 3 outlines the key characteristics of data from infectious disease outbreaks, whereas Section 4 describes common purposes for which models are related to data. Section 5 reviews some of the most important statistical methodologies and we finish with some possible future research challenges in Section 6.

2. Models for infectious disease transmission

We start by briefly reviewing some of the basic assumptions, terminology and models that are used within mathematical epidemiology. For more comprehensive and detailed introductions, see e.g. [11–13].

2.1. Terminology

Most models of infectious disease transmission are concerned with a population consisting of individuals who are potentially able to transmit the disease to one another. Here we shall focus on models that are suitable for bacterial and viral infections, as opposed to those where part of the parasite life cycle is external to the host (e.g. helminths). Models are usually defined at the level of an individual. First, individuals typically have a state that describes their health in relation to the disease in question. The standard terminology is that an individual who is able to contract the disease is called *susceptible*, one who has it but who is not yet able to infect others called *latent* or *exposed*, one who is able to pass the disease on to others called *infective*, and one who is no longer able to transmit the disease but also not susceptible is called *removed*. This final state might correspond to different things in reality depending on the disease: recovery, immunity, isolation and death are possible examples. However, the salient characteristic of removed individuals is that they do not directly affect the spread of disease. Note also the potentially confusing distinction between an *infected* individual (meaning, no longer susceptible) and an *infective* individual.

Once infected, an individual typically progresses through the latent and infective states before becoming removed. The times spent in these first two states are known as the *latent period* and *infectious period*, respectively. Models almost always assume that the progression of an individual through these states is not affected by the current states of other individuals.

Note that many models do not include specification of an individual's own response to disease, i.e. whether an individual is symptomatic, incapacitated, etc., although such detail might be introduced in a model if it is relevant to the available data (for instance, data on symptom-appearance times). Sometimes, assumptions about an individual's health might be subtly reflected through the interpretation of what the removed state corresponds to. For example, a model might assume that the appearance of symptoms coincides with removal, which in practice might mean that the individual is assumed to be too ill to carry on circulating in the population as usual. One consequence of this is that the infectious period in such models should be interpreted as an *effective* infectious period, which of course is what is important for disease transmission. Another observation is that states such as latent, infective, etc. may themselves be hard to define precisely in a clinical manner, and so such classifications are inevitably rather crude. However, they capture the essential information about individuals as regards disease transmission.

The basic states (susceptible, exposed, infective and removed) and the progression of individuals through them are frequently used in abbreviated form to describe models. For example, an SEIR model is the one in which individuals who become infected progress via the exposed and infective states to removal, with no further progression possible. Conversely an SIS model is one in which individuals can be either susceptible or infective, and moreover can become re-infected immediately at the end of the infectious period. The absence of a latent period in the SIS model is justifiable if it is used to model a disease in which the latent period is typically much shorter than the infectious period.

Finally, in specific applications it is frequently necessary to augment or adapt the basic states described above. For example, in HIV models the infectious period might be split into different sections to reflect different levels of infectivity. For some diseases, asymptomatic carriage is an important aspect of transmission and so this can be modelled explicitly. However, the basic idea that each individual is currently in one state is common to all such models.

2.2. Individual and population-level assumptions

The most basic models of disease transmission in a population assume homogeneity both in terms of individuals and the population as a whole. Here, homogeneity of individuals means that there is no variation in, for example, the propensity of individuals to become infected, their ability to pass the disease on, or the distribution of the time spent in the latent or infective states. Conversely, in many contexts such assumptions are unrealistic and likely to be important to the model behaviour, and so many models have more detailed definitions. A common example is to introduce age categories, so that infants, school children, adults and the elderly have potentially different susceptibilities.

At the population level, homogeneity means that all individuals mix together at random, which in terms of disease transmission means that a given infective is equally likely to infect any of the currently susceptible individuals. Although such an assumption might be reasonable in a small community such as a household, it rapidly becomes unrealistic as the population size increases. Some models approach this problem by dividing the population into smaller categories (e.g. age), and replacing population-wide homogeneous mixing with mixing at different rates between different categories. An alternative approach is to explicitly model population mixing structure, for example, incorporating households, schools, workplaces, transport systems, etc., so that mixing within the population occurs only within the confines of the model for population structure.

2.3. *Stochastic and deterministic models*

Models for infectious disease transmission essentially subdivide into two categories, namely deterministic models and stochastic models. The former are frequently defined via a system of ordinary or partial differential equations, which describe how the numbers or proportions of individuals in different states (susceptible, infective, etc.) evolve through time. An attractive feature of deterministic models is that it is usually fairly straightforward to obtain numerical solutions for a given set of parameter values. They are generally most effective as descriptions of reality in large populations, where, roughly speaking, laws of large numbers act to reduce the order of stochastic effects (see [12], Chapter 5 for mathematical details).

Stochastic models are usually thought of as more realistic, although their mathematical analysis is often much harder. They can capture the stochasticity seen in real-life disease outbreaks, for example, the phenomenon of fade-out in endemic diseases. They are generally defined at the level of individuals, for instance, specifying probability distributions that describe the latent or infectious periods. Stochastic models are arguably more natural than deterministic models when it comes to fitting models to data, especially for data that clearly contain inherent stochasticity.

2.4. *Transmission*

The heart of any disease transmission model is the set of assumptions made regarding transmission itself, i.e. the mechanism by which susceptible individuals become infected. The most common approach is to assume that new infections occur at a rate which is proportional to the product of the numbers of infectives and susceptibles currently in the population. Here, 'rate' has a slightly different meaning for deterministic and stochastic models. To illustrate, consider the standard deterministic SIR model for a closed population (see e.g. [14]) defined via the differential equations

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta S(t)I(t),$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta S(t)I(t) - \gamma I(t),$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I(t),$$

where $S(t)$, $I(t)$ and $R(t)$ denote the numbers of susceptible, infective and removed individuals in the population at time $t \geqslant 0$. Obviously, in the model these numbers need not be integers. A common initial condition is $(S(0), I(0), R(0)) = (N, a, 0)$, where a is typically small (e.g. $a = 1$) compared with the initial number of susceptibles N.

The standard (Markov) stochastic SIR model is defined as a continuous-time Markov chain $\{(S(t), I(t)) : t \geqslant 0\}$ with transition probabilities

$$P[(S(t+h), I(t+h)) = (s-1, i+1) | (S(t), I(t)) = (s, i)] = \beta hsi + o(h),$$

$$P[(S(t+h), I(t+h)) = (s, i-1) | (S(t), I(t)) = (s, i)] = \gamma hi + o(h),$$

these corresponding to an infection and a removal, respectively. Thus, new infections occur at the points of a non-homogeneous Poisson process (see e.g. [15, Chapter 6]) whose rate at time t is $\beta S(t)I(t)$. Note that the infectious periods of individuals in this model are independent exponential distributions with mean γ^{-1}. A commonly used variation is to instead assume that the infectious periods follow some other prescribed distribution (Gamma and constant being common examples) with associated parameters. Both the stochastic and deterministic SIR models have two parameters, namely the *infection rate* β and the *removal rate* γ. Finally, it is often the case that β is replaced by β/N when defining the above SIR models. This new parameterization is equivalent to the original for fixed N, but is appropriate if one is considering how the model behaviour changes as N varies. In particular, the β/N scaling leads to the well-known branching process approximation for the initial phase of the stochastic epidemic as $N \to \infty$ [12, Chapter 3].

3. Data

In the context of transmission modelling, it is well-known that data from outbreaks of infectious disease have two key characteristics, namely that they are usually partial and highly dependent. Here we give brief remarks about these aspects, and then describe the two most common sorts of data set that arise in practice.

3.1. Partially observed data

Data are partial in relation to transmission models whenever the actual process of transmission is not observed. This situation is almost invariably the case for human diseases, when the appearance of symptoms is usually the first sign that an individual has been infected. Even in situations where there are frequent diagnostic tests (e.g. daily swab tests in hospital studies), uncertainty about the precise moment of transmission remains. There may be other senses in which the data are partial, including cases that are not included due to under-reporting or their being asymptomatic. Another related issue is that the population of susceptibles may not be known precisely, both in the sense that some individuals might not actually be susceptible (e.g. having prior immunity due to vaccination or previous exposure to the disease), and also in the sense that individuals who enter or leave the study population may not be recorded.

In addition to these issues, at a more fundamental level there might be uncertainty about case diagnosis (e.g. an individual who is borderline in terms of the case definition being employed) and any number of issues surrounding the collection of the data themselves (e.g. inaccuracies in reporting during contact-tracing; compliance with control measures; errors in data-entry; etc.). Such issues affect any kind of statistical analysis, whether model-based or not.

3.2. Dependencies in data

Data such as daily or weekly case numbers are obviously dependent, and so any statistical analysis should take this into account. Although one can fit models containing inherent dependencies (e.g. time series models), the appeal of using transmission models is that the resulting model parameter estimates are usually meaningful in an epidemiological or biological sense. In particular, the focus of the methods described below is towards understanding the process that generated the data, as opposed to fitting the data themselves to a standard statistical model.

3.3. Temporal data and final outcome data

Although the details of data from outbreaks vary considerably from study to study, broadly speaking there are two kinds of data set, namely *temporal data* and *final outcome data*. The first is more common and typically consists of a time series of case detection times, and perhaps additional information (e.g. a time series of disease-related mortality data; covariate information such as age or sex of each case; dates of interventions, etc.). Such data are almost always aggregated, for instance, by day or week, although for most analyses it makes no material difference to try and incorporate the 'true' event times into the modelling. The second kind of data are final outcome data, which consist simply of final numbers of cases, again with covariate information. A common example is that of data collected at the level of households (e.g. [16, 17]). Obviously, final outcome data contain no explicit information about the time-course of the epidemic, and so the estimation of quantities such as the mean infectious period is not possible. However, quantities such as the basic reproduction number (see below) or relative rates of transmission between different types of individual can usually be estimated.

4. Why relate models to data?

Having defined an epidemic model, it is reasonable to perform some kind of model analysis, the aim of which is to try and understand how the model behaves under a range of initial conditions and parameter values. A typical objective of such an exercise would be to establish conditions under which the epidemic is likely to take off, or to gauge which parameters are most crucial in driving the epidemic. For serious applications, however, once the basic range of behaviour of a model is understood it is then invariably the case that the model has to be calibrated or compared somehow with data. In this section, we review some of the common situations in which models are related to data.

4.1. Evaluation of scientific hypotheses

A convenient way of addressing scientific questions is via the formulation of hypotheses which can then be tested, either formally or informally, using models and data. As an example, suppose we wish to know whether or not children are more susceptible than adults to a particular strain of influenza. One approach is to define a model that allows children and adults to have potentially different susceptibilities by using different parameters: for example, β_1 as the rate of

transmission from one infective to a susceptible child, and β_2 as the corresponding rate to a susceptible adult. Then, by fitting the model to available data one tests whether or not it is plausible that $\beta_1 = \beta_2$.

Such an approach can be used for a wide range of scientific questions (e.g. do contact precautions reduce MRSA transmission in intensive care units [18]? Were control measures effective in reducing SARS [3]?), although there are two important caveats. First, the approach relies on having sufficient data to actually address the hypothesis of interest. For example, in evaluating whether or not a new vaccine makes infectives less infectious than an existing vaccine, it is clearly necessary to have observed cases among those vaccinated. The second caveat is that any evaluation is dependent on the transmission model being used, and it is good practice to perform some kind of sensitivity analysis, perhaps by considering models with slightly different assumptions (e.g. different infectious period distributions or different infection mechanisms).

4.2. Estimation

As described above, epidemic models are usually defined from the point of view of individual-to-individual transmission events, and consequently estimates of basic model parameters often correspond directly to estimates of quantities of epidemiological interest. Common examples include the mean infectious period, the transmission rate between different types of individuals, the transmission rates from individuals infected with different strains of a virus, and the efficacy of a vaccine. Estimation frequently goes hand-in-hand with the evaluation of scientific hypotheses, since the latter tasks rely on the former.

Similarly, functions of basic model parameters may also be of interest, the most frequently seen example being the *basic reproduction number* R_0, which loosely speaking is the average number of secondary cases caused by a single infective in a large susceptible population (see e.g. [19]). For the SIR model defined above, $R_0 = N\beta/\gamma$. The basic reproduction number is important for several reasons, the salient one being that an epidemic is only likely to occur if $R_0 > 1$, and so knowledge of R_0 in turn leads to an estimate of vaccination coverage required to prevent a future outbreak in a closed population. Although we shall not consider the matter further here, it should be noted that even defining R_0 in meaningful way contains subtle difficulties when one considers more complex models (see e.g. [20]). Examples of such difficulties include defining a 'typical' single infective, or defining what is meant by a 'large' population.

4.3. What-if scenarios and prediction

A key practical benefit of epidemic modelling as a public health tool is that it can be used to evaluate different strategies for disease control or mitigation, without any need to perform field trials. Such evaluations can be both prospective (e.g. will school closures be materially effective in controlling future influenza outbreaks [21]?) or retrospective (e.g. if livestock movement restrictions had been introduced earlier, what would have been the impact on the 2001 UK foot and mouth epidemic [22]?). However, such evaluations are clearly only of value if the models in question, and the model parameters chosen, actually reflect reality. It is therefore vitally important that the models are themselves supported by data.

It should be noted that prediction usually contains two kinds of uncertainty. The first arises if the underlying epidemic model is stochastic, so that a range of realizations of the epidemic with the same underlying parameter values is possible. The second is parameter uncertainty, meaning that it can be misleading to base predictions on, for instance, maximum likelihood estimates of model parameters without taking account of the precision of such estimates. The Bayesian framework provides one way of combining these two kinds of uncertainty, see, for example [23].

4.4. Real-time data analyses

Some outbreaks of infectious disease—generally those that have the potential to cause national-level epidemics—are the subject of close monitoring and data collection. Analyses based on transmission models have a key role to play in interpreting such data in real-time (e.g. on a daily or weekly basis), and can be useful a range of activities. Examples include quantification of key aspects of the spread of disease (e.g. the serial interval, i.e. the time between successive cases in a chain of infection), assessment of ongoing containment strategies (e.g. are the current control measures working?), and prediction (e.g. how large is the outbreak likely to be?).

In practice, this kind of modelling is particularly challenging for several reasons. First, developing suitable transmission models, and writing and verifying computer code to perform statistical analyses based on such models, can be a time-consuming exercise. Second, analyses based on computationally intensive methods can sometimes take days, as opposed to hours, to actually produce results. It is therefore of benefit to develop model-based methods of analysis which are themselves efficient (see e.g. [8]). Third, there can be practical issues around converting field data into a form that is appropriate for the purposes of modelling. Finally, there can be changes to the type and quality of data collected during an outbreak, for instance, changes in case definitions, changes of what data are actually collected, etc.

5. Methods for relating models to data

As described above, a transmission model invariably contains various parameters whose values need to be defined in some manner. In the absence of data with which to estimate such parameters, it is not uncommon practice to simply assign values, these being chosen either because they appear plausible, or on the basis of previous studies in the literature. Ideally, sensitivity analyses are then performed to assess the sensitivity to the assumed values, although in practice this often becomes infeasible for more than a few parameters. At best, conclusions based on models with several parameters whose values are uncertain and not estimable from reliable data should be viewed with caution.

Conversely, if data are available with which to parameterize models, then various techniques can be used. We now briefly recall the most common methods.

5.1. Methods for deterministic models

For a given set of parameter values, a deterministic model has precisely one solution, i.e. the course of the epidemic through time is non-random. Such a model can conveniently be fitted to data using various methods, the most common of which is to minimize the sum of squares of differences between observed data and model prediction. This approach is equivalent to obtaining maximum likelihood estimates for the model in which the data are assumed to be generated by the deterministic model plus independent Gaussian errors. Our focus here is towards stochastic models, but [24] contains more details of general approaches to fitting deterministic models to data. It is also possible to adopt a Bayesian approach, see, for example, [23] for an application to epidemic modelling.

5.2. Methods for stochastic models

Given a stochastic transmission model, most inferential methods rely on likelihood. There are two common situations, namely that in which there is a tractable likelihood, and that in which there is not. These cases arise due to the interplay of the model and the data, and in particular neither is caused simply by the model nor data alone. We now explore these situations in more detail.

The first scenario, namely a tractable likelihood, tends to arise either because of various simplifying assumptions (e.g. the model assumes known fixed-length latent and infectious periods, so that partial temporal data immediately yield full temporal data), or because only final outcome data are available. In the latter case, for some transmission models it is possible to calculate, either analytically or numerically, the distribution of final outcome, and thus a likelihood can be derived. It is interesting to note that the sparse data such as final outcome data can be easier to deal with than the more detailed temporal data. Given a likelihood, inference can proceed along conventional lines, either frequentist or Bayesian inference, using tools such as maximum likelihood estimation, the Expectation Maximization (EM) algorithm, rejection sampling and MCMC methods.

The second scenario of an intractable likelihood is more common. A standard example is the case of temporal data in which (say) symptom-appearance times are observed, but infection times are not. The required likelihood can be thought of as the integral, over all possible configurations of infection times, of the joint likelihood of infection and symptom-appearance times. However, this integral is typically analytically and numerically intractable due to its high dimensionality and the non-trivial nature of the region of integration. Intractable likelihoods can also arise for final outcome data problems, as described below. We now describe four general methods for approaching the problem of an intractable likelihood.

Approximating model: Given an intractable likelihood, a natural solution is to consider a simpler approximating model under which a likelihood can be computed. One example is found in [25], in which final outcome data collected at the household level are analysed using a so-called two-level mixing model. Roughly speaking, this model considers that 'local' transmission can occur between two individuals in the same household at a rate λ_L, whereas 'global' transmission in the population at large is governed by a parameter λ_G. Thus, the fates of individuals living in different households are not independent, and in consequence it is computationally infeasible to calculate the likelihood of a given final outcome in all but the most trivial cases. The approach underlying the analysis in [25], initially developed in [26], is to replace explicit between-household interaction with a fixed probability that each individual avoids infection from outside their own household, with this probability relying on λ_G. In this way, the fates of individual households become independent, and the two-level mixing model reduces to the simpler Longini-Koopman model [27], for which a tractable likelihood exists.

A second example of model approximation, already mentioned above, is where the data consist of removal times and it is desired to fit a standard SIR model. In general, the likelihood of observing the removal times given the model parameters alone is intractable. However, if one approximates the infectious period by a fixed constant, then all the infection times are determined from the data, resulting in a tractable likelihood. Inference can then proceed along conventional lines (see e.g. [28, Section 4.4]).

Data augmentation methods: Many missing data problems within statistics are well-suited to data augmentation methods (see e.g. [29]). Suppose we have a model with parameters θ and data y. The basic idea is to introduce additional model parameters ψ, which represent missing data, in such a way that the likelihood $L(\theta, \psi) = \pi(y, \psi|\theta)$ is tractable. Inference then proceeds by estimating both θ and ψ, typically via the EM algorithm or MCMC methods.

In applying these ideas to the standard SIR model (with non-constant infectious periods) in which removals are observed, a natural choice of augmentation is to use the unobserved infection times. This approach was independently developed by Gibson and Renshaw [30] and O'Neill and Roberts [31], and has since been adapted and expanded to many other models and related kinds of data set (e.g. [9, 18, 32–35]).

Data augmentation can also be fruitfully applied to other settings where its application is less obvious. For example, Demiris and O'Neill [36] and O'Neill [37] use data augmentation methods to perform inference for the two-level mixing model mentioned above, given final outcome data in households. The method involves imputing a random graph which essentially summarizes the process of contacts between individuals. We discuss further ideas related to data augmentation below.

Martingale methods: For certain kinds of epidemic model, the so-called Martingale methods can be used for inference [12, 28]. The basic idea is to construct Martingales related to counting processes that are contained within the epidemic model (e.g. counting the number of infections by a given time), from which one then derives estimating equations for the parameters of interest. This in turn yields maximum likelihood estimators along with asymptotic results, the latter derived from the asymptotic properties of the martingales in question. The methods can be applied to temporal and final outcome data, and can also be used to perform non-parametric inference in some settings, as described in [28]. Although these methods are extremely elegant, they are also rather specialized and are not as widely applicable as most other approaches to fitting epidemic models to data.

Likelihood-free methods: It is almost always the case that stochastic epidemic models are relatively straightforward to simulate. Specifically, an actual realization of the entire course of the epidemic can usually be produced very rapidly using a computer, given a set of model parameters. This observation underlies a relatively new approach to inference called *Approximate Bayesian Computation* (ABC), which has been applied to a wide range of applications in recent years (see [24]), including inference for epidemic models with temporal data [10] and final outcome data [38].

The basic idea is as follows. Suppose we observe data y. First, propose a candidate parameter vector θ from some prior density $\pi(\theta)$. For the SIR model, θ might consist of the two basic model parameters β and γ. Next, simulate the model using θ to produce a data set x. Under the model, this has likelihood $\pi(x|\theta)$, say. Finally, compare x and y. If they are sufficiently close, say $d(x, y) < \varepsilon$ for some suitable distance function d, then accept θ; otherwise, propose a new value of θ, and so on. The output of this algorithm is a sample of model parameters drawn from the density $\pi(\theta|d(x, y) < \varepsilon)$. Clearly if d is chosen appropriately and ε is sufficiently small, then $\pi(\theta|d(x, y) < \varepsilon)$ will be a good approximation to the posterior density of interest, $\pi(\theta|y)$. In practice, the choice of d in particular is a non-trivial matter, and the subject of current research efforts.

The basic ABC algorithm described above is essentially a rejection sampling algorithm. However, it is also possible to incorporate the ABC idea into an MCMC algorithm, and into sequential Bayesian algorithms, as described in detail in [24].

6. Future challenges

We finish by briefly describing three areas that are ripe for future development. These are listed in addition to some of the areas already described above: for example, methods for real-time analysis and the application of ABC methods are also both important topics which require further research effort.

6.1. Inference for large populations and complex models

Many of the methods described above for fitting models to data work best in small-scale settings, in which it is usually possible to make effective use of models with relatively few parameters, and where the study population is not especially large. However, for large-scale settings most methods struggle due to the increased computational burden. An example is provided by data from the UK 2001 foot and mouth outbreak as considered in, for example [8, 39]. In both these papers, the authors require techniques beyond those usually employed in order to make progress with model-based analyses, the point being that standard methods struggle. The main challenge is the sheer scale of the data considered by the authors, involving thousands of farms, geographical and network information, plus data on the spread of disease itself.

A second example is found in [40] which illustrates the difficulties of fitting a standard SIR model (as opposed to, for example, time series models that incorporate some of the SIR assumptions, see [41] and references therein) to large-scale aggregated data on measles outbreaks. Here the challenge arises because of the large population size,

allied with the frequency of observation, which together make standard data augmentation methods infeasible. In some situations, methods such as non-centering parameterizations in MCMC (see e.g. [35, 37]) can be applied, although a challenge is to find generic methods to perform inference in such large-scale situations.

6.2. *Data augmentation methods*

As described above, data augmentation methods have already been used successfully for model-based analyses of different kinds of data on infectious disease outbreaks. However, since the methods are rather general in nature, there remains further scope for their novel application. Here we mention three examples to illustrate the diversity of this area. First, Cauchemez and Ferguson [40] describe a novel MCMC data augmentation method using a diffusion approximation to analyse a large-scale data set on measles outbreaks. This idea essentially makes use of an approximating model within a data augmentation framework. Second, O'Neill [37] describes data augmentation methods to analyse a sample of final outcome data from a structured population. The key idea here is to impute information on who-contacted-who during the epidemic outbreak. Third, data augmentation methods can also, in some cases, be used to provide bounds on other model parameters, the idea being to maximize or minimize over the set of possible values of the missing data. This approach is used in [42] to provide bounds for the basic model parameters and basic reproduction number in the standard SIR model.

6.3. *Model fit and model choice*

The area of model assessment for epidemic models is currently somewhat less developed than that of model fitting. In some situations standard goodness-of-fit procedures (e.g. chi-squared tests, AIC, etc.) can be used, for example household final outcome data analysed using an independent-households model (e.g. [7, 16]). However for more complex models, or for temporal data, other methods are required.

Within the setting of Bayesian inference, reversible jump MCMC methods can be used to assess the relative probability of competing models. Although such methods have been applied to epidemic models in particular settings [43, 44], a more general understanding of such an approach has yet to be developed. For example, although it is well-known within Bayesian statistics that the choice of within-model prior distribution can have a material impact on the results, this is often largely ignored in applications. Another method for assessing goodness-of-fit is to analyse residuals (see e.g. [45]). Although this approach has considerable appeal, in practice the choice of which residuals to employ can be somewhat arbitrary.

Acknowledgements

This paper is based on a presentation given by the author at the meeting 'Spatio-temporal and network modelling of diseases', Tübingen, October 2008, supported by DIMACS and ECDC. The author wishes to thank the organisors and funding providers of this event. The author also wishes to thank the Editor and two referees for helpful comments that have improved the paper.

References

1. Kim JJ, Goldie SJ. Health and economic implications of HPV vaccination in the United States. *New England Journal of Medicine* 2008; **359**:821–832.
2. Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, Ferguson N. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature* 2008; **452**:750–754.
3. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* 2004; **160**(6):509–516.
4. McBryde ES, Bergeri I, van Gemert C, Rotty J, Headley EJ, Simpson K, Lester RA, Hellard M, Fielding JE. Early transmission characteristics of influenza A(H1N1)V in Australia: Victorian state, 16 May–3 June 2009. *Eurosurveillance* 2009; **14**(42).
5. Epstein JM. Modelling to contain pandemics. *Nature* 2009; **460**:687.
6. Eubank S, Guclu H, Anil Kumar VS, Marathe MV, Srinivasan A, Toroczkai Z, Wang N. Modelling disease outbreaks in realistic urban social networks. *Nature* 2004; **429**:180–184.
7. O'Neill PD, Balding DJ, Becker NG, Eerola M, Mollison D. Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, *Series C* 2000; **49**:517–542.
8. Jewell CP, Kypraios T, Neal P, Roberts GO. Bayesian analysis for emerging infectious diseases. *Bayesian Analysis* 2009; **4**(2):191–222.
9. Lekone EP, Finkenstadt BF. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics* 2006; **62**(4):1170–1177.
10. McKinley T, Cook AR, Deardon R. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics* 2009; **5**(1). Article 24.
11. Anderson RM, May RM. *Infectious Diseases of Humans*; *Dynamics and Control*. Oxford University Press: Oxford, 1991.
12. Andersson H, Britton T. *Stochastic Epidemic Models and their Statistical Analysis*. Springer: New York, 2000.
13. Daley DJ, Gani J. *Epidemic Modelling*: *An Introduction*. Cambridge University Press: Cambridge, 1999.

14. Bailey NTJ. *The Mathematical Theory of Infectious Diseases and its Applications* (2nd edn). Griffin: London, 1975.
15. Grimmett GR, Stirzaker DS. *Probability and Random Processes* (3rd edn). Oxford University Press: Oxford, 2001.
16. van Boven M, Koopmans M, Holle MDRV, Meijer A, Klinkenberg D, Donnelly CA, Heesterbeek JAP. Detecting emerging transmissibility of avian influenza virus in human households. *PLoS Computational Biology* 2007; **3**:1394–1402.
17. Longini IM, Koopman JS, Haber M, Cotsonis GA. Statistical inference for infectious diseases: risk-specific household and community transmission parameters. *American Journal of Epidemiology* 1988; **128**:845–859.
18. Kypraios T, O'Neill PD, Huang SS, Rifas-Shiman SL, Cooper BS. Assessing the role of undetected colonization and isolation precautions in reducing methicillin-resistant staphylococcus aureus transmission in intensive care units. *BMC Infectious Diseases* 2010; **10**(29).
19. Heesterbeek JAP, Dietz K. The concept of R_0 in epidemic theory. *Statistica Neerlandica* 1996; **50**:89–110.
20. Roberts MG, Heesterbeek JAP. A new method for estimating the effort required to control an infectious disease. *Proceedings of the Royal Society of London*, *Series B* 2003; **270**:1359–1364.
21. Cauchemez S, Ferguson NM, Watchel C, Tegnell A, Saour G, Duncan B, Nicoll A. Closure of schools during an influenza pandemic. *Lancet Infectious Disease* 2009; **9**:473–481.
22. Keeling MJ, Woolhouse ME, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, Cornell SJ, Kappey J, Wilesmith J, Grenfell BT. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogenous landscape. *Science* 2001; **294**(5542):813–817.
23. Elderd B, Dukic V, Dwyer G. Uncertainty in predictions of disease spread and public-health responses to bioterrorism and emerging diseases. *Proceedings of National Academy of Sciences* 2006; **103**(42):15693–15697.
24. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* 2009; **6**:187–202.
25. Britton T, Becker NG. Estimating the immunity coverage required to prevent epidemics in a community of households. *Biostatistics* 2000; **1**:389–402.
26. Ball FG, Mollison D, Scalia-Tomba G. Epidemics with two levels of mixing. *Annals of Applied Probability* 1997; **7**:46–89.
27. Longini IM, Koopman JS. Household and community transmission parameters from final distributions of infections in households. *Biometrics* 1982; **38**:115–126.
28. Becker NG. *Analysis of Infectious Disease Data*. Chapman & Hall: London, 1989.
29. van Dyk DA, Meng X-L. The art of data augmentation. *Journal of Computational and Graphical Statistics* 2001; **10**(1):1–50.
30. Gibson GJ, Renshaw E. Estimating parameters in stochastic compartmental models. *IMA Journal of Mathematics Applied in Medicine and Biology* 1998; **15**:19–40.
31. O'Neill PD, Roberts GO. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society*, *Series A* 1999; **162**:121–129.
32. Auranen K, Arjas E, Leino T, Takala AK. Transmission of pneumococcal carriage in families: a latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association* 2000; **95**:1044–1053.
33. Streftaris G, Gibson GJ. Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling* 2004; **4**:63–75.
34. Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boëlle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine* 2004; **23**:3469–3487.
35. Neal P, Roberts GO. A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing* 2005; **15**:315–327.
36. Demiris N, O'Neill PD. Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Statistical Society*, *Series B* 2005; **67**:731–746.
37. O'Neill PD. Bayesian inference for stochastic multitype epidemics in structured populations using sample data. *Biostatistics* 2009; **10**(4):779–791.
38. Baguelin M, Newton JR, Demiris D, Daly J, Mumford JA, Wood JLN. Control of equine influenza: scenario testing using a realistic metapopulation model of spread. *Journal of the Royal Society Interface* 2010; **7**:67–79.
39. Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tidesley MJ, Savill NJ, Shaw DJ, Woolhouse MEJ. Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica* 2010; **20**:239–261.
40. Cauchemez S, Ferguson NM. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society Interface* 2008; **5**(25):885–897.
41. Morton A, Finkenstadt BF. Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*, *Series C* 2005; **54**:575–594.
42. Clancy D, O'Neill PD. Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis* 2008; **3**(4):737–758.
43. Neal P, Roberts GO. Statistical inference and model selection for the 1861 Hagelloch data set. *Biostatistics* 2004; **5**:249–261.
44. O'Neill PD, Marks PJ. Bayesian model choice and infection route modelling in an outbreak of Norovirus. *Statistics in Medicine* 2005; **24**:2011–2024.
45. Forrester ML, Pettitt AN, Gibson GJ. Bayesian inference for estimating the effectiveness of infection control measures using routine hospital data. *Biostatistics* 2007; **8**:383–401.