Statistical Science
2018, Vol. 33, No. 1, 19–33
https://doi.org/10.1214/17-STS615

Comparison and Assessment of Epidemic Models

Gavin J. Gibson, George Streftaris and David Thong

Abstract. Model criticism is a growing focus of research in stochastic epidemic modelling, following the successful addressing of model fitting and parameter estimation via powerful computationally intensive statistical methods in recent decades. In this paper, we consider a variety of stochastic representations of epidemic outbreaks, with emphasis on individual-based continuous-time models, and review the range of model comparison and assessment approaches currently applied. We highlight some of the factors that can serve to impede checking and criticism of epidemic models such as lack of replication, partial observation of processes, lack of prior knowledge on parameters in competing models, the nonnested nature of models to be compared, and computational challenges. Based on a wide selection of approaches as reported in the literature, we argue that assessment and comparison of stochastic epidemic models is complex and often, by necessity, idiosyncratic to specific applications. We particularly advocate following the advice of Box [*J. Amer. Statist. Assoc.* **71** (1976) 791–799] to be selective regarding the model inadequacies for which one tests and, moreover, to be open to the blending of classical and Bayesian ideas in epidemic model criticism, rather than adhering to a single philosophy.

Key words and phrases: Epidemic models, model comparison, model criticism, Bayesian methods, classical methods.

1. INTRODUCTION

The past three decades have seen a rapid growth in the use of stochastic models to quantify and understand the spread of infectious diseases in populations. With the advent of modern computing power and methods, which enable stochastic epidemics to be simulated efficiently and, central to this paper, to be fitted to observations of real-world epidemics, the stochastic approach is now an essential part of the modern toolkit for statistical epidemic modelling.

Key to facilitating the uptake of stochastic models, such as those described in Section 2, was the development of computational techniques for fitting stochastic models to observations within a formal Bayesian framework. The challenge arose due to the fact that observations of real-world epidemics are typically incomplete, with many aspects of the epidemic process not being directly observed. Consequently, given observations y from an epidemic, the likelihood $\pi(y|\theta)$, where θ is the model parameter vector, is often intractable. As described in Section 2, the technique of *data augmentation* [46] has proved invaluable in addressing this challenge [39].

New techniques for parameter estimation facilitated numerous studies where stochastic models were ap-

Gavin J. Gibson is a Professor of Statistics at the School of Mathematical and Computer Sciences, The Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom (e-mail: G.J.Gibson@hw.ac.uk). George Streftaris is an Associate Professor of Statistics at the School of Mathematical and Computer Sciences, The Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom (e-mail: G.Streftaris@hw.ac.uk). David Thong is a PhD Student in the School of Mathematical and Computer Sciences, The Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom (e-mail: dyt30@hw.ac.uk).

plied to characterise the dynamics of infectious diseases of humans, animals and plants. Such studies are often motivated by the need to assess the likely efficacy of putative control strategies [27, 9, 47, 11] and the fitted models are used in predictive simulation studies. In this context, questions of model assessment and comparison become particularly important.

In this paper, we discuss some challenges associated with assessment and comparison of epidemic models and review the range of approaches taken, highlighting their reported strengths and weaknesses. We find that assessment and comparison of stochastic epidemic models is a nontrivial task, the optimal approach to which depends on the practical questions that motivate the modelling. We recommend that researchers consider a wide range of approaches to model assessment and comparison and advocate pragmatism over dogmatism. We particularly recommend following the advice of Box [5] to *'worry selectively about model inadequacies'* and to target assessment methods towards detecting inadequacies that have a bearing on the underlying decision problem.

We further highlight the nonstandard nature of many inferential problems related to epidemic models, which may invalidate some standard assumptions that underlie theoretical results on model assessment. Consequently, we find that assessment of epidemic models is a field with considerable scope for innovation, and that extending the tool-kit currently available is a valuable and important pursuit for the statistical epidemic modelling community.

Modelling infectious diseases is a very broad field and we cover only some aspects in this paper. It is noted in [14] that *'model assessment is already challenging when only one source of data is involved ... and becomes even more problematic when simultaneously modelling multiple sources of information'*. Our study focuses on the former situation but we believe that its relevance extends to more complex epidemiological settings and, indeed, to fields outside epidemiology where dynamical stochastic models for partially observed, unreplicated systems are applied.

2. MODELS FOR INFECTIOUS DISEASES

Many forms of models are used to represent the dynamics of infectious diseases. Here, we focus on *stochastic* models that explicitly represent the random nature of the events during an epidemic and the lack of determinism apparent in real-world observations. This class encompasses many variates including discrete-time formulations (e.g., [16, 35, 47])—convenient for the analysis, for example, of monthly incidence data for epidemics in large populations—and continuous-time, individual-based models (e.g., [9, 21, 22]). While models of this latter type are the main focus of this paper, the messages are relevant more broadly.

Real-world epidemics of infectious diseases typically spread *via* local (as well as possibly long-range) interactions so that the population does not mix homogeneously. For this reason, epidemic models often represent spatial aspects of spread explicitly and it is often these aspects which demand most scrutiny in model assessment exercises [41].

We mainly focus on epidemic models exemplified by the spatio-temporal SEIR model in which a population is partitioned into susceptible (S), exposed (E), infectious (I) and removed (R) classes through which (in the absence of controls measures) individuals move sequentially. It is often assumed that the infection process, governing transitions from S to E, is Markovian. In the general case, infection spreads through contact between susceptible and infectious individuals (*secondary* infection) and additionally may occur *via* environmental or other external sources (*primary* infection). Under these assumptions, individual i, susceptible at time t, becomes exposed during $[t, t+dt]$ with probability

$$p = \left(\varepsilon + \beta \sum_{j \in I(t)} K(d_{ij}, \alpha)\right) dt + o(dt), \tag{1}$$

where ε and β, respectively, quantify the rate of primary and secondary infection.

The summation in (1) represents the combined challenges from hosts infectious at time t. The function $K(d_{ij}, \alpha)$, or *spatial kernel*, where α is a parameter and d_{ij} represents the distance between i and infectious host j, describes how the challenge presented to i by j declines with distance. The model is completed on specifying distributions for the random times $T_E \sim \pi_{\theta_E}$, and $T_I \sim \pi_{\theta_I}$ spent by an individual in the E and I classes, respectively, where θ_E and θ_I are additional parameters. Here, T_E represents the time taken by an exposed host to incubate the disease to the stage where they can infect others, while T_I represents the time taken following the onset of infectiousness before removal which, depending on the epidemic, may occur through the acquisition of immunity, death, hospitalisation or other means. Choices for these distributions include the Gamma distribution [38, 22] or the Weibull distribution [44, 45, 28].

This generic framework can be extended, for example, to represent further routes of infection or multilevel mixing patterns. Where the model is used to represent a meta-population [9, 21, 22, 47] it may be extended to represent the dependence of susceptibility or infectivity of each subpopulation on the local species mix [22]. Alternative compartmental structures can be proposed, with class I comprising asymptomatic and symptomatic subcompartments [41]. Moreover, formulation (1) can incorporate a wide range of patterns of heterogeneous mixing, since d_{ij} need not represent Euclidean distance, but may take discrete values, for example, dependent on whether i and j are in the same household or classroom [26, 36].

2.1 Bayesian Model Fitting

When fitting epidemic models to observations the partial nature of data y means that certain transitions between compartments are not directly observable. A common scenario arises when only removal events ($I \to R$) are observed (e.g., [39, 7]). In other cases, observations may capture the status of individuals at discrete, and possibly sparse, survey times—yielding a series 'snapshots' of the population state. This latter scenario arises frequently in studies of diseases of arboreal pathogens such as that reported in [37]. In epidemics in metapopulations of households, y may record numbers of infections in each household [26] rather than precise times of infections. Data on monthly numbers of reported data (e.g., [35]) may be generated through some random sampling process applied to the (unobserved) infected population.

These scenarios involve data y that are a censored, filtered or noisy version of a notional 'complete' process and parameter estimation demands that challenges of 'missing data' are routinely faced. This motivates the adoption of the Bayesian approach, with 'missing' data accommodated using data augmentation as follows. Let θ denote the parameter vector of an epidemic model (and include components parameterising the observation process as appropriate) and, for the moment, let x denote the times and nature of all events in the epidemic—the 'complete data'. Then the likelihood $\pi(x|\theta)$ is typically computationally tractable.

Where the observed y is an incomplete or 'noisy' version of x, then we may conveniently express $\pi(x, y|\theta) \propto \pi(x|\theta_1)\pi(y|x, \theta_2)$ where θ_1 and θ_2 parameterise the epidemic and observation processes, respectively. In other settings such as the 'snapshot' scenario or removal-only setting, where $y = g(x)$, then $\pi(x, y|\theta) = \pi(x|\theta)$ for $y = g(x)$ and zero otherwise. In either scenario given y only, the desired likelihood is

$$\pi(y|\theta) \propto \int \pi(x, y|\theta)\, dx, \tag{2}$$

which is often intractable.

The Bayesian approach avoids the need to compute (2) by assigning a prior parameter distribution, $\pi(\theta)$, and then sampling from the *joint* posterior density $\pi(x, \theta|y)$—most commonly using Markov chain Monte Carlo methods (MCMC). When data y do not distinguish between individuals in certain states, the number of unobserved event times in x may not be fixed so that the state space of the Markov chain comprises components with differing dimension. This arises for removal data where the number of infected individuals, and hence the number of unobserved infection times, by the end of the observation period is unknown. For this reason, 'reversible-jump' methods [20], used to design chains with multi-dimensional state spaces, are frequently used. The literature includes numerous accounts of this approach in action (e.g., [39, 18, 17, 9]). While x usually specifies the temporal history of the epidemic as regards the timing of transitions undergone by individuals, it may additionally include details of infection trees ('who-infected-whom'). A topic of major current importance in epidemic modelling is the integration of phylogenetic information on pathogens with other epidemic data in this inferential framework (see, e.g., [49, 23, 34, 29]).

Data augmentation has been described as a 'gold standard' [30] for parameter estimation for epidemic models, and is feasible for analysis of large-scale epidemics involving many thousands of hosts. Effective algorithms can also be designed using particle filtering algorithms [25]. Others have applied Approximate Bayesian Computation (ABC) to epidemic models [4, 30], replacing y by summary statistics $T(y)$ and effectively estimating $\pi(\theta|T(y))$ through forward simulation of the model. There are nevertheless benefits arising from the use of data augmentation when it comes to model criticism—with several techniques in later sections exploiting its capacity to impute latent processes.

While availability of parameter estimation tools increases the potential of stochastic modelling to inform the design of control measures for epidemics, it also increases the demand for model assessment tools to underpin the validity of any decisions taken using stochastic models.

Assessing the fit of epidemic models is not straightforward. When observing real-world epidemics, there

is often a *lack of replication*, with only a single realisation of an epidemic process observed. While some epidemics on extensive, spatially distributed populations, may be considered as a 'patchwork' of replicates of a single process, environmental heterogeneities may render the assumption that the same model applies to all regions invalid [41]. In the Bayesian framework, this lack of replication means that parameter posterior densities may be highly nonnormal, complicating in turn the use of model assessment methods such as the Deviance Information Criterion (DIC, [43]) due to the diverse choice of point estimators for θ (see Section 3.4) and the nonquadratic nature of the log-likelihood.

The partial nature of observations both motivates and complicates the use of Bayesian techniques. Although missing data are naturally accommodated using the approach, posterior distributions can be very sensitive to the choice of prior distributions, even for the simplest epidemic models. Suppose that each host in a (large) population of size N is subjected to an infectious challenge of size α/N starting from time $t = 0$ when the population is fully susceptible, each infected individual remaining so for a period drawn from $\text{Exp}(\mu)$ until removal. Only removal events are observed. The system follows an immigration-death process with infections arising as a Poisson process with rate α with subsequent removal rate μ. The removal events can be shown to follow an inhomogeneous Poisson process with time-varying rate

$$R(t) = \alpha\big(1 - e^{-\mu t}\big).$$

In [18] the problem of inferring μ given observations $y = (t_1, t_2, \ldots, t_n)$, the first n removal times, was considered. When α is assigned an $\text{Exp}(\gamma)$ prior, it was shown that the marginal posterior density of μ collapses to unit mass at 0 as as $\gamma \to 0$, a phenomenon which can be attributed to unbounded level curves for the likelihood function. In contrast, the likelihood function when infections and removals are observed is very well behaved. Similar problems exist when assigning vague priors to parameters governing unobserved transitions for more realistic models such as the SIR and SEIR. This sensitivity of the posterior to the prior effectively means that measures of model fit that depend on the parameter posterior (such as the DIC or the use of posterior predictive checks) may likewise be sensitive to the choice of prior. Moreover, as discussed in Section 3, even when the parameter posterior distribution for a given model is insensitive to the prior, some measures used for model comparison such as Bayes factors should nevertheless be expected to exhibit sensitivity to the prior.

Given these complexities, it is understandable that a wide range of approaches are taken to assess the fit of epidemic models. We review some of these in Section 3.

3. APPROACHES TO MODEL CHOICE AND ASSESSMENT

Approaches to comparison and assessment for epidemic models draw on a range of statistical philosophies in determining quantitative measures of model fit. Methods range from the purely Bayesian, which expresses all uncertainties in the form of probability distributions updated in the light of data using the laws of conditional probability, to frequentist approaches based on benchmarking aspects of observed data against sampling properties of parameterised models. On this spectrum, we may immediately locate Bayesian model choice and the use of Bayes factors [36, 26] at one extreme. On the other hand, assessment of time-series models based on comparing spectral properties of data with simulated spectra using maximum-likelihood parameter estimates (see [35]) would lie towards the opposite (frequentist) extreme. We may also distinguish approaches that compare one model with specified alternatives (model comparison) from those that test model adequacy in the absence of any explicitly specified alternative. In practice, approaches used to assess epidemic models often blend Bayesian and frequentist thinking and may represent alternative models with various degrees of specification.

3.1 Bayesian Model Choice and the Use of Bayes Factors

Despite the widespread adoption of the Bayesian approach for estimating parameters there have been comparatively few attempts to pursue a fully Bayesian approach to epidemic model comparison. Suppose that we have a set of competing epidemic models $M_1, \ldots, M_k$ with parameter vectors $\theta_1, \ldots, \theta_k$, equipped with prior distributions $\pi_j(\theta_j)$, $j = 1, \ldots, k$ and that these cover the range of models that could govern an observed epidemic. Following a purely Bayesian approach [15], a prior probability p_j is assigned to each model and, given data y, the model posterior distribution can be defined as

$$\Pr(M_j|y) \propto p_j \Pr(y|M_j) = p_j \int \pi_j(y|\theta_j)\pi(\theta_j)\, d\theta_j.$$

In theory, the model posterior distribution can be investigated using reversible-jump MCMC (RJMCMC) techniques [20] to construct a Markov chain whose state space is the union of the parameter spaces of the competing models and whose stationary distribution is the posterior distribution of the 'parameter' of this *expanded* model. Posterior model probabilities are then obtained from the chain's relative long-term frequencies of occupancy of the parameter spaces of the competing models.

However, even when the likelihood functions $\pi_j(y|\theta_j)$ are tractable, implementation of RJMCMC for model comparison may be difficult. As described in Section 2, it is often necessary to use RJMCMC when fitting a single model j to partial data. In the model-comparison setting, each model parameter vector θ_j may have to be augmented with hidden data components appropriate to that model, x_j, and the space of the augmented parameter $\theta_j^* = (\theta_j, x_j)$ for any single model may already be a union of components of differing dimension. Competing models may not share the same compartmental structure, for example, when comparing an SEIR formulation with the simpler SIR formulation with observed removal times only [18]. The former model implies unobserved transitions $S \to E$ and $E \to I$ while the latter implies unobserved transitions from $S \to I$ only. This means that the nature of x_j may vary over models. Nevertheless, some epidemic modellers have taken on the computational challenges of Bayesian model choice and demonstrated its feasibility in certain scenarios.

In [36], Bayes factors and posterior model probabilities for models of the 1861 Hagelloch Measles epidemic are computed. This is a rich data set, including *inter alia* for each case the time of appearance of different symptoms, but neither the infection time nor the removal time. In the most complex model, the force of infection presented by an infective i to a susceptible j was modelled as

$$\alpha_{ij} = \beta_H \delta_{H(i)H(j)} + \beta_C \delta_{C(i)C(j)} + \beta_G \exp(-\theta d_{ij}),$$

where δ denotes the Kronecker delta symbol, $H(i)$ and $C(i)$ respectively indicate the household and classroom of individual i and d_{ij} denotes the geographical distance between i and j, with θ governing the dependence of the transmission rate on distance. Also, β_H, β_C denote within-household and within-classroom contact rates, respectively, while β_G gives the global contact rate. In [36], the importance of these effects was assessed by comparing the full model with three variants obtained by excluding one of the above effects (distance, household and classroom) using RJMCMC. For all models in the comparison, the nature of the missing data—infection and removal times—is the same. Therefore, the RJMCMC only requires to formulate dimension-changing moves related to the model parameters θ_j as opposed to the augmenting variables x_j, easing its implementation. Runs with simulated datasets found that, even though in general the correct model was selected, model ranking was affected by the choice of the priors $\pi(\theta_j)$, with more informative priors yielding more posterior support for the the full model in some datasets. For epidemics generated from models with no household effect, there was difficulty in identifying the correct model, due to an element of confounding of the household and spatial effects β_G. Prior specification had influence over model ranking in this latter situation. Models with alternative spatial kernel functions were considered but not in a Bayesian comparison; it was merely noted that the ranking of the four model variants (via posterior model probabilities) was relatively robust to the choice of spatial kernel and also to the replacement of imputed infection times with fixed estimates.

Bayesian model comparison is also carried out in [26] for final outcome data for epidemics in metapopulations of households, where the data record the numbers of susceptible and removed individuals in each household at the conclusion of the epidemic. Again the focus of this study is the force of infection and the contribution from local and global contacts. In [26], it is assumed that an individual makes global infectious contacts with the population according to a Poisson process with rate λ_G and additionally with members of its household according to an additional independent Poisson process with rate $n\lambda_L$ where n is household size. For both local and global contacts, the infectee is selected uniformly from the relevant population. Models are compared using simulated data and data from the Tecumseh study of influenza [33].

The paper compares three model variants:

- M_1: with two parameters λ_G and λ_L;
- M_2: $\lambda_G = \lambda_L = \lambda$;
- M_3: $\lambda_L = 0$.

As in [36], RJMCMC is used as the main approach to model comparison. However, in [26] the augmenting data, x_j, consist of information on the number of global and local infectious contacts and the corresponding recipients of the contacts; given these data posterior distributions of the Poisson rates become tractable.

Moves between models in the algorithm involve only $\theta = (\lambda_G, \lambda_L)$ rather than components of the augmenting data. Models are compared in a pairwise fashion and the methods are used to estimate the Bayes' factors associated with each comparison.

The results of [26] highlight the sensitivity of the Bayes factor to the choice of prior. As the priors on the rates λ_G and λ_L become increasingly vague, when there is even a single individual who escapes infection, the Bayes' factor will increasingly favour the simpler model. In this case, the Bayes' factor is very sensitive to the data, changing dramatically if data describing a completely infected population is modified to include a single susceptible at the end of the epidemic.

In summary, studies that have applied Bayesian model comparison to epidemic models have found that the computational challenges can be overcome to some extent. Reversible-jump methods can be applied given sufficient ingenuity in designing the moves, although this may nevertheless be challenging if competing models do not share a common latent process x that can augment each model parameter vector θ_j. Although not discussed in detail here, we note that alternative approaches for computing Bayes' factors can be considered. For example, path-sampling is also applied in [26], rejection sampling in [10] and there has been recent progress in using the method of power posteriors [1] to compute marginal likelihoods.

However, even when the computational difficulties can be overcome, the fundamental sensitivities of the conclusions of Bayesian model choice to the prior distributions of the competing models remain. When informative priors for model parameters can be used—for example, if an emerging epidemic is believed to be governed by one of several, well understood processes—the approach might be feasible. In settings where little prior information is available, then conclusions of Bayesian model comparisons should be qualified in the light of the approach's tendency to penalise both the complexity of model j, as measured by the dimension of θ_j, and ignorance regarding θ_j.

3.2 Posterior Predictive P-Values and Checks

Box [6] advocates that parameter estimation should be seen as a process which is best pursued within a Bayesian framework while *model criticism* requires consideration of sampling properties of models. It is therefore appealing to consider assessment methods that combine Bayesian and frequentist ideas in the way that the posterior predictive P-value (PPP-value) [31] does.

The posterior predictive P-value is calculated as follows. Given data y and a model specified by $\pi(y|\theta)$, with prior $\pi(\theta)$, we consider the posterior predictive distribution of the data in a replicate experiment, y^{rep}, or more usually some function of the observations, $T(y^{\text{rep}})$, and we compute, a PPP-value as

$$\begin{aligned} p(y) &= \Pr\bigl(T\bigl(y^{\text{rep}}\bigr) > T(y)|y\bigr) \\ &= \int \Pr\bigl(T\bigl(y^{\text{rep}}\bigr) > T(y)|\theta\bigr)\pi(\theta|y)\,d\theta. \end{aligned}$$

The quantity $p(y)$ can be interpreted as the posterior probability of a more extreme value of T in the next replicate experiment. Small (or large) values of $p(y)$ therefore indicate that the observed data are extreme and provide evidence against the modelling assumptions. The use of PPP-values is conservative since the prior predictive distribution of $p(y)$ is stochastically less variable than a $U(0,1)$, as demonstrated in [31]. In effect, this implies the existence of α_0 such that the quantile (with respect to this prior predictive distribution) of any observed $p(y) < \alpha_0$ is strictly less than $p(y)$. Since the calculation of PPP-values is based on the posterior distribution $\pi(\theta|y)$, then the results are only sensitive to the prior distribution $\pi(\theta)$ to the extent that the posterior $\pi(\theta|y)$ is sensitive to the prior.

Many researchers have used the general notion of posterior predictive checking, whereby one or more test statistics $T(y)$ is examined for departures from its posterior predictive distribution. Because of the complexity of epidemic data sets, particularly spatio-temporal ones, it is often necessary to use a range of test statistics to capture spatial and temporal characteristics. Commonly chosen summary statistics include disease progress curves where

$$T(y) = \bigl(I(t_1), I(t_2), \ldots, I(t_k)\bigr)$$

summarises the number of infected individuals in the population at observation times $(t_1, \ldots, t_k)$. Comparison of the observed $T(y)$ with its predictive distribution is then done by benchmarking $T(y)$ against an envelope of progress curves drawn from the predictive distribution. An example of posterior predictive checking using disease progress curves is found in [41] where a range of spatio-temporal models for the spread of Huanglongbing (HLB) virus in citrus orchards are fitted to an extensive spatio-temporal data set. A range of models are compared on the basis of posterior predictive distributions of disease progress curves and also, reflecting the explicit spatio-temporal nature of the data set, measures of spatial correlations in infected sets—in this case a modified Moran index

for presence/absence with a ring weighting function. By evaluating the index for a range of choices of radii in the weighting function, a correlation function $C(d)$ can be generated; see [41] for details.

A related approach is taken in [37] which analysed epidemics of citrus canker in an urban landscape, testing models using the predictive distribution of disease progress curves and measures of spatial autocorrelation. In particular, [37] makes use of a spatial autocorrelation function defined as

$$C_T(d) = \frac{\rho_{II}(d,T) - \rho_I(T)}{\rho_I(T)(1-\rho_I(t))}$$

where $\rho_{II}(d,T)$ denotes the proportion of pairs of hosts at distance d apart who are both infected at time T and $\rho_I(T)$ denotes the proportion of hosts infected at time T. This function is estimated from data in [37] using the nonparametric spine correlogram [3] and provides a richer description of spatial structure than any spatial correlation index.

The use of disease progress curves for posterior predictive checking is only possible when available data allow the infected set to be specified at certain observation times, as would be the case with 'snapshot' data. It would however not be appropriate in scenarios where data record only removal times, for example. For such situations, suitable statistics include that of [7] where an SIR model with change-points in parameters is fitted to removal-time data on a smallpox epidemic in a village, originally considered in [2], and a richer data set on an outbreak of a respiratory disease. Posterior predictive checking was done using the time of the k^{th} removal for various values of k. This choice of test statistic is natural given that the study aimed to elicit evidence of temporal heterogeneity in model parameters. Alternatively, test statistics based on cumulative numbers of removals at specified observation times $T(y) = (R(t_1), \ldots, R(t_n))$ could be defined when only removal data are available.

Posterior predictive checking presents one instance where Box's advice—to worry selectively about misspecification—can be applied in selecting appropriate test statistics. If a model is to be used to predict future incidence of a disease, then its ability to capture a disease progress curve would be a meaningful indicator of utility. On the other hand, if a model is to be used to target control measures at individual hosts, then its ability to predict a disease progress curve may be of little value if it has little power to predict the status of individual hosts correctly. One approach that stresses prediction of individual status is the suite of accuracy measures used to assess spatio-temporal models from [24] for the 2001 FMD epidemic in [47]. Here, a point estimate for the model parameter vector, obtained from the entire epidemic, is used to generate epidemics on spatially distributed host populations comprising UK farm locations, taking as their initial conditions the status of farms at an early point t_0 in that epidemic. Measures of model accuracy are then derived from the distribution of the proportion of farms whose state is predicted correctly at some future time t_1, where the measure can be chosen to be specific to a particular state, or subset of the host population partitioned according to region. These distributions are benchmarked against the distribution of the corresponding proportion when the real data are replaced by an independent simulation from the model, which captures the inherent *repeatability* of the model. Although in [47] the parameter estimation framework is not Bayesian, the methods readily extend to a Bayesian analysis on replacing point estimates of parameters with draws from $\pi(\theta|y)$ in the simulation of measures. Through the suitable choice of t_0 and t_1, long-term or short-term predictive accuracy can be assessed depending, for example, to the particular timescales on which control measures could be deployed.

It is clear that posterior predictive checking has proved useful as a means of excluding models on the basis of their inability to reproduce key aspects of observed epidemics. Although a PPP-value carries evidence against a model without specifying an alternative model, the choice of summary statistics can be motivated by some prior hypotheses regarding the model deficiencies or by the importance of certain deficiencies for the purpose of the modelling. Spatial correlation measures are natural candidates to test spatio-temporal models where the concern may lie with the validity of the choice of spatial kernel function, while measures of accuracy of prediction of host state [47] are clearly relevant if the model is to inform targeted control strategies.

On the other hand, the reliance on low-dimensional summary statistics for complex data sets may reduce the power of posterior predictive checking to detect lack of fit, much as the use of summary statistics in Approximate Bayesian Computation [30, 4] inflates the variance of parameter posterior distributions. In the following section, we consider developments of posterior predictive checking that potentially may provide more sensitive tests of inadequacy.

3.3 Latent Residual Tests and Noncentred Parameterisations

An extension of the notion of PPP-values has been used by some authors to tailor tests of model adequacy

to particular forms of misspecification. Regarding PPP-values, we note the following:

1. Following [31], we can replace the statistic $T(y)$ with a function $T(y,\theta)$, known as a *discrepancy variable*, and define the PPP-value as

$$p(y) = \Pr\big(T\big(y^{\text{rep}},\theta\big) > T(y,\theta)|y\big) = \int \Pr\big(T\big(y^{\text{rep}},\theta\big) > T(y,\theta)|\theta\big)\pi(\theta|y)\,d\theta.$$

For example, when y is a random sample of size n from $N(\mu,\sigma^2)$ then a natural choice of discrepancy variable is $\frac{(\bar{y}-\mu)^2}{\sqrt{\sigma^2/n}}$.

2. When defining discrepancy variables or test statistics, we could consider T as being dependent on an unobserved latent process x imputed in a Bayesian analysis, rather than on y only, formulating a more general PPP-value as

$$p(y) = \Pr\big(T\big(x^{\text{rep}},\theta\big) > T(x,\theta)|y\big) = \int \Pr\big(T\big(x^{\text{rep}},\theta\big) > T(x,\theta)|\theta\big)\pi(\theta,x|y)\,d\theta.$$

We note that $p(y)$ is the posterior expectation of the quantity

$$\Pr\big(T\big(x^{\text{rep}},\theta\big) > T(x,\theta)|\theta\big) = p(x,T,\theta),$$

which has a *frequentist* interpretation as a classical P-value. Since the posterior distribution of this classical P-value gives a more complete summary of evidence against the assumed model than does its expectation, we prefer to focus on $\pi(p(x,T,\theta)|y)$ as the object of interest in what follows. Of course, results may be reported in terms of summaries of this distribution such as its expectation or $\pi(p(x,T,\theta) < \alpha|y)$.

This distribution $\pi(p(x,T,\theta)|y)$ can be interpreted as the the posterior belief of Bayesian (B) regarding the P-value calculated by a classical observer (C) who observes y and x and tests a simple null hypothesis under which the system obeys the assumed model parameterised by θ. The identification of these separate entities B and C resonates with ideas in [19] where the metaphor of Freud's theory of personality is used to characterise hybrid statistical logic. In our particular adoption of this metaphor, we consider the *id* to represent our instinctive appetite for representing complex processes using parsimonious but effective models. The *ego* (B, above) uses Bayesian reasoning to transform the simplified model and beliefs into a framework for inference and prediction. The *superego* (C above) plays the role of the conscience, using classical reasoning to benchmark observed quantities against their sampling distributions to assess the validity of B's assumptions. In this framework, it is C who specifies the test whose result B should impute, using their suspicions regarding model misspecification to specify x and T, prior to B observing y.

An important principle in this kind of hybrid reasoning is the following. Suppose that $\pi_j(y,x_j|\theta)$, $j = 1,\ldots,k$ represent models for the joint distribution of (y,x_j) all of which specify the same *marginal* model $\pi(y|\theta)$ and share a common prior distribution $\pi(\theta)$. Then observation of y alone carries no information on the relative validity of these models. Put another way, y carries exactly the same evidence *against* every model with marginal $\pi(y|\theta)$. We therefore have complete freedom in the choice of the latent process x.

We can design tests for epidemic models using this approach using the idea of functional-model [12] representations (cf. noncentred parameterisations [40]). We design a functional-model representation of a stochastic model $\pi(x|\theta)$ by identifying a stochastic process r with known distribution independent of θ, and a function $h_\theta()$ such that $x = h_\theta(r) \sim \pi(x|\theta)$. A simple example of a functional model is the generation of a continuous random variable *via* inversion of the distribution function, where $r \sim U(0,1)$ represents the quantile of the generated value $x = F_\theta^{-1}(r)$.

In the epidemic context, we can construct a functional model for the complete epidemic data x (times and nature of all events). Then, given data y, and the functional model $x = h_\theta(r)$, we investigate the posterior distribution $\pi(\theta,r|y)$ and impute the result of a classical test carried out on r, to test its compliance with its known sampling distribution. This approach was adopted in [17] to test spatio-temporal SI models for the spread of *R solani* in populations of radish, where r was chosen to be the set of Sellke thresholds [42] of individual hosts, where the Sellke threshold of individual j, r_j, represents a notional threshold through which the infection time of individual j is specified as the instant at which the integrated rate of infection undergone by j reaches r_j. The Markovian infection process is obtained on choosing the r_j to be i.i.d. Exp(1). Given 'snapshots' of the infected set at discrete times the authors of [17] imputed the results of a Kolmogorov–Smirnov test of compliance r with the Exp(1), eliciting evidence of lack of fit from the posterior distributions of the imputed P-values, which was attributed to heterogeneity over replicate experiments.

A related approach was taken in [22] to assess models for the 2001 FMD epidemic. These authors imputed infectious periods for farms [assumed in the

model to be Gamma$(4, \beta)$] and compared the distribution of the imputed infectious period, scaled by β, to a Gamma(4, 1) distribution. They did not formally compute the posterior distribution of a P-value but, rather, considered an ensemble of simulated values generated during the course of an MCMC analysis, which they compared to the Gamma(4, 1), finding no evidence against the assumed model.

In spatio-temporal epidemic modelling, the form of the kernel function $K(d, \theta)$ is more often of interest given its importance for designing ring-culling strategies. In [28], a functional-model for a general spatio-temporal SEIR model is considered in which the process r is composed of four independent i.i.d. $U(0, 1)$ processes, r_1, r_2, r_3, r_4. Under the mapping $x = h_\theta(r_1, r_2, r_3, r_4)$, the time of each subsequent infection event is determined from the process r_1, with $-\log r_1$ defining a sequence of 'population-level' Sellke thresholds. Processes r_3 and r_4 specify the quantiles of the sojourn periods in the E and I class, respectively, for each infection as it occurs, while r_2 determines the particular I-S pair responsible for each infection event. The key device in designing a test sensitive to misspecification of spatial kernel is the manner of selecting the particular I-S pair to determine each new infection. In [28], given the time of the j^{th} infection, t_j, the link is selected by first assigning every I-S link a weight equal to rate of infection across the link. Links are then ordered according to decreasing weight. The particular link is selected by considering the cumulative sum of the ordered links and identifying the particular link responsible for this cumulative sum reaching the value $r_{2j}W$ where W denotes the sum of the weights. The joint posterior $\pi(\theta, r_1, r_2, r_3, r_4|y)$ is explored. If the kernel function K has been misspecified (e.g., by underestimating the propensity for long-range transmission by assuming an exponentially bounded form when a power-law relation is more appropriate), then when the process r_2 is imputed, some systematic deviation from a $U(0, 1)$ should be anticipated. In [28] the P-values were imputed for an Anderson–Darling test applied to r_2 to demonstrate that the approach can detect kernel misspecification in simulated data sets. The approach was applied to compare alternative spatial kernels for the spread of giant hogweed throughout the UK, using 'snapshot' data on the distribution of the species at three time points. Significant evidence was found against models with long range spatial interactions and which omitted the effect of habitat suitability. Tests can readily be designed using the approach for other forms of misspecification, for example, the erroneous assumption of an isotropic kernel, using alternative orderings of I-S links in the definition of the functional model. Viewed in this framework, the model assessment methods of [22] effectively apply tests to the imputed process r_3, to detect misspecification of the latent period.

It should be expected that the evidence against the model elicited using a latent-residual test will be sensitive to the particular choice of functional model and the latent processes imputed. For example, infection-link residuals could be defined via functional models that use alternative orderings of the infection links to that of [28], where links were ordered from largest to smallest, but the resulting tests might not exhibit the same sensitivity to misspecification of the tail properties of the infection kernel. We emphasise the importance of taking account, where possible, of the anticipated modes of misspecification in the design of latent-residual tests.

The extension of posterior predictive checking to latent processes offers a number of advantages. By exploiting the fact that the augmenting data can be specified in any way we wish, so long as the marginal model $\pi(y|\theta)$ is preserved, it offers a way of tailoring the imputed tests to detect suspected 'axes' of misspecification, without the need to formulate fully an alternative model structure. However, there are potential disadvantages. Since imputation of the latent process conditions on the model, the assumptions of the latter are reflected in the imputed data. If these are not sufficiently constrained by the observations y, then the approach cannot detect misspecification. Reducing to the absurd, we see that if the test were applied to a latent process r, *independent* of y given θ, then the posterior (and prior) distribution of the imputed P-value (assuming a continuous test statistic) would be $U(0, 1)$. While the use of a summary statistic $T(y)$ to calculate a PPP-value may lose power due to discarding information, then imputing a latent process to test may lose power due to reinforcement of the model being tested.

3.4 Latent Likelihood-Ratio Tests

The desire to compare specific model alternatives while avoiding difficulties of Bayesian model comparison (e.g., prior selection within models, Lindley–Bartlett paradoxes, chain mixing) has also motivated the development of latent likelihood-ratio (LLR) tests for comparing a given, assumed model with a specified alternative within an otherwise Bayesian framework (e.g., [44, 45]). Let the assumed model be M_1, with parameter vector θ_1 whose prior is $\pi_1(\theta_1)$, and suppose that the hypothesised alternative is M_2 with

unknown parameter θ_2. We simulate the posterior distribution of a likelihood ratio statistic, and associated P-value, calculated from an imputed latent process x for which $\pi_1(x|\theta_1)$ and $\pi_2(x|\theta_2)$ are tractable, where π_1 and π_2 denote distributions of x under M_1 and M_2, respectively. Specifically, we fit M_1 using data augmentation to sample from $\pi(\theta_1, x|y)$ and impute the values of

$$\Lambda(x, \theta_1) = \frac{\pi_1(x|\theta_1)}{\pi_2(x|\hat{\theta}_2(x))},$$

where $\hat{\theta}_2(x)$ denotes a point estimate (*e.g.*, the MLE) of θ_2 calculated from the imputed x. Only M_1 is fitted using Bayesian methods, avoiding the difficulties associated with the fully Bayesian approach. Therefore, when M_2 is the more complex model, it is not required to fit it in the Bayesian framework, as would be the case when using the Deviance Information Criterion (discussed in the next section). Repeated simulation of x under M_1 parameterised by the imputed θ_1, or asymptotic approximations based on the χ^2 distribution when the models are nested, can then provide the empirical sampling distribution of $\Lambda(x, \theta_1^{(k)})$ under M_1 and consequently a posterior tail probability for the latent likelihood-ratio statistic, $\Pr(\Lambda(x, \theta_1^{(k)}) \leq \Lambda(x^{(k)}, \theta_1^{(k)})|\theta_1^{(k)})$, where k denotes the MCMC iteration, to evaluate evidence against model M_1 when compared to M_2.

We note the assumption that the latent process x is common to both M_1 and M_2 so that latent likelihood ratio tests are most readily implemented in settings where M_1 and M_2 share the same compartmental structure and allowable transitions but differ in the representation of the parametric processes governing the transitions. We note that, when the compartmental structures differ, for example, when comparing SIR and SEIR formulations, it may nevertheless be possible (see, e.g., [18]) to construct a common latent process.

The method was used in [45] to compare SEIR models where the infection process under the Sellke construction is modelled using either commonly employed Exp(1) (M_1) stochastic thresholds of tolerance to infection, or a more general alternative with Weibull distributed thresholds (M_2). Using simulated epidemic outbreak data, it demonstrated the power of the LLR test for selecting the correct model. The approach was also used in [44] to impute results of an ANOVA test applied to viraemia measures on sheep partitioned according to depth in an imputed infection tree, in order to elicit evidence of a 'passage' effect in 2 experimental epidemics of FMD. In this case, M_1 was the null model under which expected viraemic levels did not vary with depth in the infection tree, with M_2 allowing variation with depth. The posterior distributions of P-values obtained did not suggest any strong evidence of the passage effect.

As with the methods of Section 3.3, the ability of the approach to distinguish between models may be limited when a large volume of information, represented by x above, is imputed using the assumed model. Nevertheless, this approach offers an easily implementable means of comparing specified alternative models without the need to introduce additional complexity into a Bayesian analysis.

Finally, we contrast the latent likelihood ratio test with the Bayesian approach of using the Savage–Dickey ratio and generalisations [48] to effect model comparisons by estimating Bayes' factors. Both approaches share the common feature of not having to obtain the posterior density for both competing models. The latter approach relies on using simulations from the posterior distribution of the parameter vector of the more complex model in a nested setting, in contrast to the latent likelihood ratio approach which can be implemented by Bayesian fitting of the less complex model in any comparison and is not restricted to nested settings.

3.5 Deviance Information Criterion

The deviance information criterion was introduced in [43], as a measure of model fit that can be applied in the Bayesian framework to assess the fit of a given model. It has a close relationship to Akaike's Information Criterion (AIC) and approximates this in sufficiently regular, large-sample settings when the log-posterior is quadratic. In its original form, given observations y, the DIC is computed as

$$DIC_1 = -4E_\theta\{\log \pi(y|\theta)|y\} + 2\log \pi(y|\tilde{\theta}), \tag{3}$$

where $\tilde{\theta}$ denotes an estimate of θ such as the posterior mean or mode. It can be written as

$$DIC_1 = -2E_\theta\{\log \pi(y|\theta)|y\} + P_D,$$

where $P_D = -2E_\theta\{\log \pi(y|\theta)|y\} + 2\log \pi(y|\tilde{\theta})$ measures the effective number of parameters in the model. The DIC defined in (3) is an example of an *observed* DIC as only the data that are actually observed enter into its calculation. Competing models are assessed by calculating and comparing their DICs, the model with the smallest DIC being preferred.

So long as $\pi(y|\theta)$ is tractable, an observed DIC can be calculated as a straightforward addendum to a

Bayesian analysis using MCMC, and the DIC is often seen as a Bayesian technique. Arguably, the use of the DIC is not truly Bayesian in spirit. Model comparison is made on the basis of the difference of the DIC between models, but this difference comprises terms each of which requires conditioning on a different model for the data. It is therefore hard to interpret *differences* in any DIC from the perspective of any single Bayesian observer, in the way that a PPP-value has an interpretation for a Bayesian as a posterior probability. For this reason, we see DIC as less Bayesian than the latter approach and have deferred its treatment to this point.

There are several instances of the DIC being used to assess epidemic models. In [13], a range of variants of the observed DIC, differing in the manner in which P_D is defined, are compared in a simulation study carried out using a discrete-time spatio-temporal epidemic model with latent susceptible classes. The study focusses on how effective the variants are in favouring the true model and how this changes with the amount of information available on the latent classes. They find considerable variation in the ability of the DICs to 'select' the correct model, with one particular variant (DIC_3 from [8]) proving the most robust to lack of information on latent classes.

In epidemiological settings, the intractability of $\pi(y|\theta)$ means that DIC_1 defined in (3) (and other observed DICs) cannot be calculated. In [8], the authors present a range of alternative, 'missing-data' formulations that can be used in cases where $\pi(y|\theta)$ is intractable but $\pi(y,x|\theta)$ is tractable, for some latent process x. These are natural candidates for comparing epidemic models fitted using data augmented MCMC. However, as there exist an infinity of ways of extending $\pi(y|\theta)$ to a 'missing data' model $\pi(y,x|\theta)$, there are an infinite number of 'missing data' DICs. Moreover, for any given choice of x, [8] offers several alternative definitions of a DIC.

Particular versions from [8] that have been applied in epidemic modelling include

$$\begin{aligned} DIC_4 &= -4E_{\theta,x}\{\log\pi(y,x|\theta)|y\} \\ &\quad + 2E_x\{\log\pi(y,x|E_\theta(\theta|y,x)|y\}. \end{aligned} \tag{4}$$

A second variant used in epidemic modelling is

$$\begin{aligned} DIC_6 &= -4E_{\theta,x}\{\log\pi(y,x|\theta)|y\} \\ &\quad + 2E_x\{\log\pi\big(y,x|\hat\theta(y)\big)|y\}, \end{aligned} \tag{5}$$

where $\hat\theta(y)$ is an estimate of θ derived from the posterior $\pi(\theta|y)$. Also applied is

$$\begin{aligned} DIC_8 &= -4E_{\theta,x}\{\log\pi(y|x,\theta)|y\} \\ &\quad + 2E_x\{\log\pi(y|x,\hat\theta(y,x))|y\}, \end{aligned} \tag{6}$$

where $\hat\theta(y,x)$ is now an estimate of θ derived from the posterior $\pi(\theta|y,x)$.

DIC_4 was used in [26] to compare models with multiple levels of mixing, as an alternative to Bayesian model comparison. In particular, [26] suggested DIC_4 generally gave a model ranking which was more stable than that obtained using Bayes' factors with respect to increasing vagueness of the parameter priors, although this was not the case for all the models considered. In [28], DIC_4 and DIC_8 were used to compare spatio-temporal models with differing spatial kernels and latent period distributions using simulated and real epidemic data. It was found that the ranking of competing models could vary with the choice of DIC.

Given this diversity of rankings, we should consider how the choice of DIC may be motivated in any particular scenario. Central to this question is the notion of 'focus' of the inference, discussed in [43] and [32], which relates to those aspects of a model most relevant to the purpose of the modelling study. Embedding a marginal model $\pi(y|\theta)$ in a joint model $\pi(y,x|\theta)$ and using the latter to compute a 'missing-data' DIC may change the inferential focus from that intended, as explained below. This contrasts with the use of data augmentation as a tool to facilitate simulation from $\pi(\theta|y)$, whose nature is invariant with respect to the choice of x.

In Section 3.3, we motivated the design of infection-link residuals from the perspective of a notional observer of a latent process. We can view missing data DICs in a similar manner by expressing particular forms of the DIC as the posterior expectation of a measure calculated by a latent observer. The focus (and appropriateness) of any particular DIC measure may then be characterised from the perspective of this observer. For example, DIC_4 can be expressed as

$$\begin{aligned} DIC_4 &= E_x\big[-4E_\theta\{\log\pi(y,x|\theta)|y,x\} \\ &\quad + 2\log\pi(y,x|E_\theta(\theta|y,x)|y\big]. \end{aligned}$$

This corresponds to the posterior expectation of the value of DIC_1 computed by an observer of x in addition to y, making DIC_4 a 'natural' generalisation of DIC_1. In contrast,

$$\begin{aligned} DIC_6 &= E_x\big[-4E_\theta\{\log\pi(y,x|\theta)|y,x\} \\ &\quad + 2\log\pi\big(y,x|\hat\theta(y)\big)|y\big]. \end{aligned}$$

This latent observer of (y,x) in this case appears less rational than the first as their estimate $\theta(y)$ depends on y only, and the additional information provided by x is ignored.

Displaying DIC_8 in a similar manner we see that

$$DIC_8 = E_x\big[-4E_\theta\{\log\pi(y|x,\theta)|y,x\} + 2\log\pi\big(y|x,\hat{\theta}(y,x)\big)|y\big].$$

Here, our latent observer of x estimates θ from x and y but considers a partial likelihood $\pi(y|x,\theta)$, rather than the full likelihood $\pi(x|\theta)\pi(y|x,\theta)$, in their calculations. The focus of this observer therefore lies with the model for y conditional on x.

In epidemic modelling, DIC_8 may sometimes be straightforward to compute, for example, when $\pi(x, y|\theta)$ combines a latent dynamical model for x coupled with an observation model for y given x, so that

$$\pi(x, y|\theta) = \pi(x|\theta_1)\pi(y|x,\theta_2),$$

where θ_1 and θ_2 parameterise the dynamical and observational models, respectively. However, should the intended focus lie with the correctness of $\pi(x|\theta_1)$, it may not be be the most appropriate choice.

In [28], DIC_8 was computed for the case of a colonisation process for the spread of Giant Hogweed in the UK, governed by a spatio-temporal SI model, to compare models with alternative spatial kernels, where x represented precise (unobserved) colonisation times of sites in a metapopulation and y represented detections at distinct survey times. Here, θ_2 is the probability that a site is reported as colonised given that it is colonised while θ_1 comprises the parameters in the SI model. In the light of the above discussion of DIC_8 it could be argued that this variant of the DIC was not the most appropriate given that the focus of interest in [28] was the choice of spatial kernel. Indeed, if our focus is on the dynamics of x rather than the observational model, we can motivate a further missing-data DIC defined by

$$DIC^* = E_x\big[-4E_{\theta_1}\{\log\pi(x|\theta_1)|x\} + 2\log\pi(x|E_{\theta_1}(\theta_1|x))|y\big],$$

this being the posterior expectation of DIC_4 as calculated by an observer whose focus lay with $\pi(x|\theta_1)$.

While some variants of the DIC may be readily computed for competing models fitted separately by Bayesian data augmentation, avoiding the complexity of full Bayesian model comparison, it remains unclear how much weight to place on the rankings obtained. The typically nonnormal nature of parameter posteriors means there is no unique estimate of θ to use in a DIC calculation, and due consideration should be given to which missing-data formulations may be most appropriate in the light of the aims of the modelling exercise.

4. CONCLUSIONS

From our survey, we have found that model choice for stochastic epidemic models is far from simple and that no single approach fits all scenarios. The issues of lack of replication, intractable likelihoods, partial observation, prior ignorance, coupled with the fact that competing models may have very different, nonnested structures contribute to the challenge. At the same time, model choice is often of paramount importance in epidemiological studies where predictions of the likely efficacy of control strategies can be highly sensitive to the model.

There is a wide range of approaches to model choice and assessment for epidemic models and this diversity presents a challenge in itself, as different approaches may lead to different rankings of models in any comparison. All the approaches and measures we have outlined have advantages and disadvantages in terms of *inter alia* their sensitivity to choices of prior, their ease of interpretation and the ease of implementation in the epidemiological context. All are potentially valuable tools in model assessment. We attempt to summarise some general principles that might guide the epidemic modeller in their choice of approaches.

1. It is important to recognise that all models are simplistic, imperfect representations of reality which may nevertheless be potentially useful. It is therefore important that the purpose of the modelling is made clear at the outset. If the model is to be used as a 'lens', for example, to determine whether a given infection rate might be nonstationary over time, or a particular effect should be included in the model, then using different models may yield the same qualitative conclusions when fitted to data and differences in the quality of fit as determined by measures described here may not be of major importance. However, if a model is to be used in predictive mode (often the case in epidemics) then quantitative differences in the predictive distributions of key outcomes over competing models assume major significance.

2. Following from (1), once the purpose of the modelling is agreed, then, as advocated by Box [5], we recommend that modellers should be selective regarding the particular misspecifications for which they aim to test. For example, if the purpose is to design controls for the spread of an epidemic in space and time using ring-culling, then the spatial kernel function is an important aspect of the model, while the precise distributional form of the latent period (particularly if it

is small with respect to the timescales over which infections occur) may be of little importance. If a model is used to assess the effectiveness of control based on quarantining infected individuals, then it may be important that the distribution of the infectious period is properly specified in the model.

3. Statistical measures of model fit, in the light of (2), should be selected and tailored to be sensitive (and specific) to important modes of misspecification. This consideration should guide, for example, the particular choice of missing-data DIC (Section 3.5), the particular latent residual process imputed (Section 3.3), the model forms selected for comparison (Sections 3.1 and 3.4) or the summary statistics used for posterior predictive checking (Section 3.2).

Following principles (1)–(3) can help ensure that models are ranked in any comparative exercise according to their potential to inform practically important decisions. The different approaches we have outlined all have strengths and weaknesses with respect to this targeted strategy.

In the case of the Bayesian model choice and the use of Bayes factors, one can ensure that practically important model alternatives are included in the space of models. However, the experience of epidemic modellers with Bayesian model comparison (e.g., [26]) suggests that, even when the computational difficulties of carrying out Bayesian model comparison can be overcome, the sensitivity of the approach to within-model priors would make this approach challenging unless informative priors can be identified.

Turning to the use of PPP-values as a means of model assessment, the challenge here is to identify summary statistics and discrepancy measures that are sensitive to important modes of misspecification. While disease progress curves and measures of spatial autocorrelation provide natural candidates, it is important to note that the reduction of a complex dynamic to a few summary statistics discards much information. We argue that the extension of the posterior predictive checks and classical tests to imputed processes (as in Sections 3.3 and 3.4) has potential in epidemic modelling, with constructions such as infection link residuals being particularly valuable for testing for kernel misspecification, but note the tendency of the imputation process to reinforce the modelling assumptions, reducing the power of the approach to detect misspecification.

The DIC, calculation of which can be easily embedded in data-augmentation MCMC algorithms, offers a convenient measure of model fit with demonstrable power to detect the 'true' models in simulation studies. However, with its numerous alternative forms, particularly in the 'missing-data' setting, care must be taken to preserve the 'focus' of the assessment on the most practically important aspects of models.

Throughout we have focussed on continuous-time models but the approaches considered can in principle be extended to discrete-time formulations (e.g., [13, 16, 35, 47]). Model selection methods including posterior predictive checks and the DIC have already been applied extensively in the discrete-time setting [13]. The construction of latent likelihood ratio tests, described in Section 3.4, should be readily achievable for discrete-time models as there are no systematic barriers to performing the inferences of latent processes and the subsequent likelihood computations in the discrete-time setting. However, we note that the construction of functional-model representations to design latent residual tests, such as those based on infection-link residuals discussed in Section 3.3, requires some thought for discrete-time models as the strict ordering in the infection times, which simplifies the construction of the residual processes r_1, r_2, r_3, r_4 in Section 3.3, is lost when infections are constrained to occur at discrete times t_i, say, with the potential for multiple infections to occur at the same time. The design of a *parsimonious* residual process, analogous to r_2, that controls the selection of the subset of susceptible hosts that become infected at t_i, conditioned on the state of the system at t_{i-1} appears to be a nontrivial challenge worthy of investigation. We note further that some of the approaches discussed may prove challenging if Approximate Bayesian Computation (ABC) [4, 30] is the preferred means of model fitting since this approach actively eschews the use of imputed, latent processes in favour of summaries of the observed data. While posterior predictive checking should fit naturally with the use of ABC we should not expect the methods discussed in Sections 3.3–3.5, many of which rely explicitly on imputed processes, to be immediately applicable when ABC is used.

Our final conclusion is that no technique offers a panacea for for model criticism in epidemic modelling and that, arguably, the most effective approach is to draw on different philosophies to tailor techniques to the situation at hand. We find that, so long as the complexity of the model being fitted is suitably constrained, then the Bayesian approach appears best suited to the estimation of parameters, while classical approaches offer flexibility to explore model deficiencies in order

to inform what should be seen as as continuous process of model criticism and refinement as advocated by Box [5].

ACKNOWLEDGEMENTS

The authors would like to thank the referee and Associate Editor for their constructive comments which have helped to improve the paper substantially. David Thong received financial support from a Doctoral Training Account from the Engineering and Physical Sciences Research Council.

REFERENCES

[1] ALHARTHI, M. (2015). Bayesian model assessment for stochastic epidemic models, Ph.D. thesis, Univ. Nottingham.

[2] BAILEY, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*, 2nd ed. Hafner Press, New York. MR0452809

[3] BJØRNSTAD, O. N. and FALCK, W. (2001). Nonparametric spatial covariance functions: Estimation and testing. *Environ. Ecol. Stat.* **8** 53–70. MR1844501

[4] BLUM, M. and TRAN, V. (2010). HIV with contact tracing: A case study in approximate Bayesian computation. *Biostatistics* **11** 644–660.

[5] BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

[6] BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

[7] BOYS, R. J. and GILES, P. R. (2007). Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates. *J. Math. Biol.* **55** 223–247. MR2322850

[8] CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673. MR2282197

[9] CHIS STER, I., SINGH, B. K. and FERGUSON, N. M. (2009). Epidemiological inference for partially observed epidemics: The example of the 2001 foot and mouth epidemic in Great Britain. *Epidemics* **1** 21–34.

[10] CLANCY, D. and O'NEILL, P. D. (2007). Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scand. J. Stat.* **34** 259–274. MR2346639

[11] COOK, A. R., GIBSON, G. J., GOTTWALD, T. and GILLIGAN, C. A. (2008). Constructing the effect of alternative intervention strategies in historic epidemics. *J. R. Soc. Interface* **5** 1203–1213.

[12] DAWID, A. P. and STONE, M. (1982). The functional-model basis of fiducial inference. *Ann. Statist.* **10** 1054–1074. MR0673643

[13] DEETH, L., DEARDON, R. and GILLIS, D. (2015). Model choice using the Deviance Information Criterion for latent conditional individual-level models of infectious disease spread. *Epidemiologic Methods* **4** 47–68.

[14] DE ANGELIS, D., PRESANIS, A. M., BIRRELL, P. J., TOMBA, G. S. and HOUSE, T. (2015). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* **10** 83–87.

[15] DRAPER, D. (1995). Assessment and propagation of model uncertainty. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 45–97. MR1325378

[16] FINKENSTÄDT, B. F. and GRENFELL, B. T. (2000). Time series modelling of childhood diseases: A dynamical systems approach. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 187–205. MR1821321

[17] GIBSON, G. J., OTTEN, W., FILIPE, J. A. N., COOK, A., MARION, G. and GILLIGAN, C. A. (2006). Bayesian estimation for percolation models of disease spread in plant populations. *Stat. Comput.* **16** 391–402. MR2297539

[18] GIBSON, G. J. and RENSHAW, E. (2001). Likelihood estimation for stochastic compartmental models using Markov chain methods. *Stat. Comput.* **11** 347–358. MR1863505

[19] GIGERENZER, G. (1993). The superego, the ego, and the id in statistical reasoning. In *A Handbook for Data Analysis in the Behavioral Sciences*: *Methodological Issues* (G. Keren and C. Lewis, eds.). Lawrence Erlbaum Associates, Hillsdale, NJ.

[20] GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

[21] JEWELL, C. P., KEELING, M. J. and ROBERTS, G. O. (2008). Predicting undetected infections during the 2007 foot and mouth disease outbreak. *J. R. Soc. Interface* **6** 1145–1151.

[22] JEWELL, C. P., KYPRAIOS, T., NEAL, P. and ROBERTS, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Anal.* **4** 465–496. MR2551042

[23] JOMBART, T., DIDELOT, X., CAUCHEMEZ, S., VIBOUD, F. C. and FERGUSON, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10** e1003457.

[24] KEELING, M. J., WOOLHOUSE, M. E. J., SHAW, D. J., MATTHEWS, L., CHASE-TOPPING, M., HAYDON, D. T., CORNELL, S. J., KAPPEY, J., WILESMITH, J. and GRENFELL, B. T. (2001). Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic dispersal in a heterogeneous landscape. *Science* **294** 813–817.

[25] KING, A. A., IONIDES, E. L., PASCUAL, M. and BOUMA, M. J. (2008). Inapparent infections and cholera dynamics. *Nature* **454** 877–880.

[26] KNOCK, E. S. and O'NEILL, P. D. (2014). Bayesian model choice for epidemic models with two levels of mixing. *Biostatistics* **15** 46–59.

[27] LAU, M. S. Y., DALZIEL, B. D., FUNK, S., MCCLELLAND, A., TIFFANY, A., RILEY, S., METCALF, C., JESSICA, E. and GRENFELL, B. T. (2017). Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. *Proc. Natl. Acad. Sci. USA* **114** 2337–2342.

[28] LAU, M. S. Y., MARION, G., STREFTARIS, G. and GIBSON, G. J. (2014). New model diagnostics for spatio-temporal systems in ecology and epidemiology. *J. R. Soc. Interface* **11** 20131093. DOI:10.1098/rsif.2013.1093.

[29] LAU, M. S. Y., MARION, G., STREFTARIS, G. and GIBSON, G. J. (2015). A systematic Bayesian integration of epidemiological and genetic data. *PLoS Comput. Biol.* **11**. e1004633. DOI:10.1371/journal.pcbi.1004633.

[30] MCKINLEY, T., COOK, A. R. and DEARDON, R. (2009). Inference in epidemic models without likelihoods. *Int. J. Biostat.* **5** Art. 24, 39. MR2533810

[31] MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160. MR1311969

[32] MENG, X.-L. and VAIDA, F. (2006). Comment on article by Celeux et al. [MR2282197]. *Bayesian Anal.* **1** 687–698. MR2282201

[33] MONTO, A. S., KOOPMAN, J. S. and LONGINI, I. M. (1985). Tecumseh study of illness: XIII, influenza infection and disease, 1976–1981. *Amer. J. Epidemiol.* **121** 811–822.

[34] MORELLI, M. J., THÉBAUD, G., CHADŒUF, J., KING, D. P., HAYDON, D. T. and SOUBEYRAND, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.* **8** e1002768, 14. MR3007333

[35] MORTON, A. and FINKENSTÄDT, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 575–594. MR2137255

[36] NEAL, P. J. and ROBERTS, G. O. Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics* **5** 249–261.

[37] NERI, F. M., COOK, A. R., GIBSON, G. J., GOTTWALD, T. R. and GILLIGAN, C. A. (2014). Bayesian analysis for inference of an emerging epidemic: Citrus canker in urban landscapes. *PLoS Comput. Biol.* **10** e1003587. DOI:10.1371/journal.pcbi.1003587.

[38] O'NEILL, P. D. and BECKER, N. G. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics* **2** 99–108.

[39] O'NEILL, P. D. and ROBERTS, G. O. (1999). Bayesian inference for partially observed epidemics. *J. R. Stat. Soc., A* **162** 121–129.

[40] PAPASPILIOPOULOS, O., ROBERTS, G. O. and SKÖLD, M. (2007). A general framework for the parametrization of hierarchical models. *Statist. Sci.* **22** 59–73. MR2408661

[41] PARRY, M., GIBSON, G. J., PARNELL, S., GOTTWALD, T. R., IREY, M. S., GAST, T. C. and GILLIGAN, C. A. (2014). Bayesian inference for an emerging arboreal epidemic in the presence of control. *Proc. Natl. Acad. Sci. USA* **111** 6258–6262.

[42] SELLKE, T. (1983). On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Probab.* **20** 390–394. MR0698541

[43] SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380

[44] STREFTARIS, G. and GIBSON, G. J. (2004). Bayesian analysis of experimental epidemics of foot-and-mouth disease. *Proc. R. Soc. Lond., B Biol. Sci.* **271** 1111–1117.

[45] STREFTARIS, G. and GIBSON, G. J. (2012). Non-exponential tolerance to infection in epidemic systems—modelling, inference and assessment. *Biostatistics* **13** 580–593.

[46] TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

[47] TILDESLEY, M. J., DEARDON, R., SAVILL, N. J., BESSELL, P. R., BROOKS, S. P., WOOLHOUSE, M. E. J., GRENFELL, B. T. and KEELING, M. J. (2008). Accuracy of models for the 2001 foot-and-mouth epidemic. *Proc. R. Soc. Lond., B Biol. Sci.* **275** 1459–1468.

[48] VERDINELLI, I. and WASSERMAN, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *J. Amer. Statist. Assoc.* **90** 614–618. MR1340514

[49] YPMA, R., BATAILLE, A., STEGEMAN, A., KOCH, G., WALLINGA, J. and VAN BALLEGOOIJEN, W. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. Lond., B Biol. Sci.* **279** 444–450.