Submitted to Statistical Science

Evidence synthesis for stochastic epidemic models

Paul J Birrell, Daniela De Angelis and Anne M Presanis

MRC Biostatistics Unit, University of Cambridge

Abstract. In recent years the role of epidemic models in informing public health policies has progressively grown. Models have become increasingly realistic and more complex, requiring the use of multiple data sources to estimate all quantities of interest. This review summarises the different types of stochastic epidemic models that use evidence synthesis and highlights current challenges.

Key words and phrases: evidence synthesis, state-space models, epidemic modelling, mechanistic modelling.

1. BACKGROUND

Epidemic models have become increasingly central to public health decision making, providing quantitative support to the efficient planning of health-care resources, the determination of optimal control strategies and the assessment of interventions to interrupt disease transmission. All of these require knowledge on hidden aspects of epidemics, such as current disease prevalence, severity, incidence and transmission, which can only be indirectly inferred through modelling. As a consequence of this crucial role of models, the methodologies underpinning epidemic modelling have come under increasing scrutiny. This has lead to more frequent adoption of rigorous approaches to linking models to data [21], increasing realism and therefore model complexity, and the need to use rich data arrays to guarantee reliable estimation. The result has been a recent proliferation of models incorporating data from multiple sources [e.g. 1, 13].

We will summarise and review some selected key examples in this literature by characterising models using a common construct. Most epidemic processes can be expressed through a state vector $\boldsymbol{X}_t$ representing unobservable characteristics of the epidemic and a vector of observable quantities $\boldsymbol{Y}_t$, under a generalised parameter-driven state-space framework:

$$\boldsymbol{X}_t|\boldsymbol{X}_{1:t-1}, \boldsymbol{Y}_{1:t-1} \sim p_{\boldsymbol{\phi}}\left(\cdot|\boldsymbol{X}_{t-1}\right) \qquad \text{(state equation)} \tag{1.1}$$

$$\boldsymbol{Y}_t|\boldsymbol{X}_{1:t}, \boldsymbol{Y}_{1:t-1} \sim p_{(\boldsymbol{\phi},\boldsymbol{\eta})}\left(\cdot|\boldsymbol{X}_t\right) \qquad \text{(observation equation)} \tag{1.2}$$

Paul Birrell is a Senior Investigator Statistician, Daniela De Angelis (e-mail: daniela.deangelis@mrc-bsu.cam.ac.uk) is a Programme Leader and Anne Presanis is a Senior Investigator Statistician all at the MRC Biostatistics Unit, University of Cambridge, School of Clinical Medicine, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK.

where $t = 1, \ldots, T$ and the $p(\cdot|\cdot)$ are appropriately chosen probability density functions [10]. Equation (1.1) governs the development of the epidemic system, characterised by a vector of parameters $\boldsymbol{\phi}$. Equation (1.2) relates the underlying epidemic process to relevant potential data $\boldsymbol{Y}_t$. These data are typically imperfect observations associated with $\boldsymbol{X}_t$, constrained by the limitations of surveillance schemes and subject to (a vector of) nuisance parameters, $\boldsymbol{\eta}$. State vectors consist of all latent quantities that may change over time, usually probabilistically, and $\boldsymbol{\phi}$ governs their temporal development. In some cases, the state vector is simply a deterministic function of $\boldsymbol{\phi}$. More commonly, epidemic models are compartmental, partitioning a population according to, for example, infection status. The distribution of individuals in each model compartment is part of the state vector, as is any quantity describing model dynamics that evolves over time, *e.g.* incidence of infection λ_t [6] or the transmission potential β_t, the disease transmission rate conditional on contact between an infectious and a susceptible individual [34].

The focus of the statistical analysis could be to estimate unobserved system states $\boldsymbol{X}_{1:T}$ either sequentially (filtering) or retrospectively (smoothing), and/or to make inference about components of $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\eta})$ that have some crucial interpretation. These parameter components might measure some headline statistic for the epidemic, such as the epidemic's reproductive number R_0, the average number of secondary infections caused by a single primary infection in a wholly susceptible population, or the effect of an intervention. This inference, ideally, would be based on direct observations $\boldsymbol{Y}_t$ on the states $\boldsymbol{X}_t$, *i.e.*

$$\boldsymbol{Y}_t = \boldsymbol{X}_t + \boldsymbol{\eta}^T \boldsymbol{\epsilon}_{Y,t}, \text{ where } \boldsymbol{\epsilon}_{Y,t} \sim N(\boldsymbol{0}, \boldsymbol{I}). \tag{1.3}$$

However equation (1.3) implies observation of, for instance, new infections as they occur, which, especially in large populations, is rarely feasible. More realistically, data are indirectly related to the quantities of interest and inference becomes possible only through the integration of data from multiple sources. Thus, given $\boldsymbol{\theta}$, $\boldsymbol{Y}_t = (\boldsymbol{Y}_t^1, \ldots, \boldsymbol{Y}_t^N)$ is a collection of N independent data sources with observed values $\boldsymbol{y}_t = (\boldsymbol{y}_t^1, \ldots, \boldsymbol{y}_t^N)$.

Evidence does not just come in the form of data. There are also modelling assumptions that underlie the parametric forms of $p_{\boldsymbol{\phi}}(\cdot)$ and $p_{(\boldsymbol{\phi},\boldsymbol{\eta})}(\cdot)$, based on relevant literature, expert opinion and/or collateral data not included in the model. In particular, pragmatic choices might need to be made over which parameter components can realistically be estimated by the available data, and which components it is prudent to assume to be known from literature (but can be varied as part of a sensitivity analysis). Synthesis of these kinds of evidence can be formalised by adopting a Bayesian framework centered on the posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{x}_{1:T}|\boldsymbol{y}_{1:T}) \propto p(\boldsymbol{y}_{1:T}|\boldsymbol{x}_{1:T}, \boldsymbol{\theta})\, p(\boldsymbol{x}_{1:T}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{1.4}$$

where $p(\boldsymbol{\theta})$, the prior distribution for $\boldsymbol{\theta}$, encodes all that is known of $\boldsymbol{\theta}$ from sources external to the present study. The posterior distribution represents a natural synthesis of this additional external information with $\boldsymbol{y}_{1:T}$.

In this paper, we shall provide an overview of evidence syntheses in stochastic epidemic modelling where multiple types of data are explicitly used in an integrated analysis. In Section 2 we will focus on non-mechanistic statistical models

for epidemic data, i.e. where transmission is not explicitly modelled. Initially these models will be static, and the aim of the analysis is to estimate the current state of an epidemic. This set-up will then be extended by adding a time dimension, initially to estimate time-varying disease incidence. In Section 3 we consider how multiple sources of data are used for inference in mechanistic models for disease transmission. In Section 3.1, the dynamics governing transmission are assumed to be deterministic (*i.e.* $\text{var}\,(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) = 0, \forall t$), so that stochasticity is only provided by the observational component (1.2). Section 3.2 reviews evidence syntheses in epidemic models with stochastic dynamics (*i.e.* $\text{var}\,(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) \neq 0$). The paper concludes with a discussion, identifying some ongoing and future challenges in the use of multiple datasets in stochastic epidemic modelling.

2. NON-MECHANISTIC EPIDEMIC MODELLING

2.1 Static Models

Often estimation of the state of an epidemic at a particular point in time is of interest. In such examples, static or "snapshot" models are used, and the temporal evolution in equations (1.1) and (1.2) is not relevant:

$$\begin{aligned}\boldsymbol{X} &\sim p_{\boldsymbol{\phi}}(\cdot)\\ \boldsymbol{Y} &\sim p_{\boldsymbol{\theta}}(\cdot|\boldsymbol{X})\,.\end{aligned}$$

In many cases, $\boldsymbol{X}$ will be a deterministic function of $\boldsymbol{\phi}$, *i.e.* $\boldsymbol{X} \equiv \boldsymbol{X}(\boldsymbol{\phi})$, or can be integrated out of the analysis entirely if estimation of $\boldsymbol{\phi}$ is the focus. We shall therefore write $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{X})$.

As anticipated in Section 1, data come in the form of N independent components $\boldsymbol{y} = (\boldsymbol{y}^1, \ldots, \boldsymbol{y}^N)$, where each $\boldsymbol{y}^n, n \in 1, \ldots, N$ may be multivariate. The aim of the evidence synthesis is to estimate a set of K *basic* parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ from the complete array of information. Each dataset $\boldsymbol{y}^n$ is assumed to inform a function $\psi_n = \psi_n(\boldsymbol{\theta})$ of the basic parameters, where ψ_n is denoted a *functional* parameter. If $\psi_n(\boldsymbol{\theta}) \equiv \theta_k$, the data $\boldsymbol{y}^n$ are said to *directly* inform θ_k, whereas if the function is more complex and/or a function of multiple components of $\boldsymbol{\theta}$, $\boldsymbol{y}^n$ *indirectly* informs one or more parameters. Denote by $\boldsymbol{\psi}$ the collection of functional parameters $(\psi_1, \ldots \psi_N)$ informed by $\boldsymbol{y}$. Assuming conditional independence of each dataset, the likelihood is then

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{n=1}^{N} L_n(\psi_n(\boldsymbol{\theta}); \boldsymbol{y}^n)$$

where each $L_n(\psi_n(\boldsymbol{\theta}); \boldsymbol{y}^n)$ is the contribution of $\boldsymbol{y}^n$ to the basic parameters. Either this likelihood is maximised, in a frequentist setting, or, in the Bayesian setting we consider here, a posterior distribution is obtained (equation (1.4)), summarising all information, both direct and indirect, as well as prior, on the basic parameters.

Such an evidence synthesis model can be represented as a directed acyclic graph (DAG) that encodes the conditional independence assumptions [25]. In the example of Figure 1, each basic parameter $\theta_k \in \boldsymbol{\theta}$, denoted by double circles, is a *founder* node of the DAG, *i.e.* using family relationships to describe the relationships between nodes, it has no parents, only descendants. Functional parameters $\psi_n \in \boldsymbol{\psi}$ (single circles) are children of the basic parameters of which

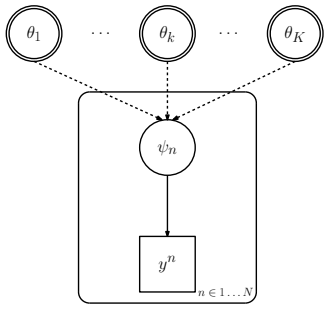


FIG 1. *Directed Acyclic Graph of a model with basic parameters, functional parameters and data.*

they are functions, with the dashed arrows denoting the (deterministic) functional relationship. By contrast, a solid arrow denotes a distributional (stochastic) relationship between nodes. Squares denote observed quantities $\boldsymbol{y}^n$. In a more complex hierarchical model with multiple levels, *consequential* nodes internal to the DAG may be either deterministically or stochastically related to their ancestors or descendants. Repetition over variables is represented by 'plates', rounded rectangles surrounding the repeated nodes, as for example the repetition of each $\boldsymbol{y}^n, n \in 1 \ldots N$ informing a different functional parameter ψ_n in the figure.

Evidence synthesis methods in the context of healthcare were introduced in a synthesis of HIV prevalence data from different groups, reviewed in [1]. These have inspired a proliferation of comprehensive evidence syntheses for static models of infectious diseases, including Hepatitis C virus [e.g. 28], influenza severity [e.g. 35] and campylobacter infection [2]. A key example is the estimation of HIV prevalence, undiagnosed prevalence in particular, in different European countries [11, 14, 31], including annually for the United Kingdom (UK) (`https://www.gov.uk/government/statistics/hiv-in-the-united-kingdom`). Estimates are produced from multiple routine HIV surveillance datasets combined with contemporaneous cross-sectional survey data.

Figure 2(a) presents a DAG of this general approach, summarised in [14]. Here the $\boldsymbol{\psi}$ are expressed as a function of basic parameters $\boldsymbol{\theta} = \{(\rho_g, \pi_g, \delta_g) : g \in 1, \ldots, G\}$, where ρ_g is the proportion of a population in a particular risk group g for HIV; π_g is the proportion of group g infected; and δ_g is the proportion of infections in group g that are detected (diagnosed). Example functional parameters include $\psi_{ng}(\boldsymbol{\theta}) = \pi_g(1 - \delta_g)$, the prevalence of undiagnosed infection, and $\psi_{mg}(\boldsymbol{\theta}) = N\rho_g\pi_g\delta_g$, the number of diagnosed infections in group g. As the data are either proportions or counts, the likelihood is comprised of binomial and Poisson terms whose parameters are the functional parameters $\boldsymbol{\psi}$. Two key challenges in building such an evidence synthesis are: sparse data leading to identifiability issues and requiring hierarchical models to borrow strength, i.e. extending the DAG of Figure 2(a) vertically; and in contrast, multiple data sources informing the same parameter, with a resultant potential for these data to conflict. Such conflicts are typically due to unaccounted biases, and need to be detected, measured and resolved [see 13, and references therein].

The motivation behind evidence synthesis is to frame all the available information on the state of an epidemic within a single integrated analysis to address identifiability. For a number of reasons, however, including computational effi-

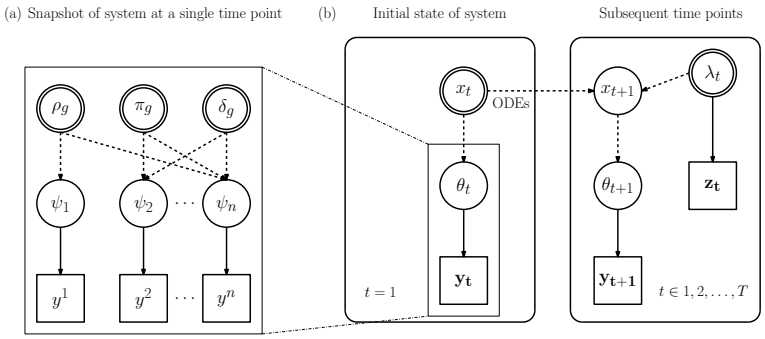


FIG 2. *(a) DAG of a HIV prevalence model with basic parameters* $\boldsymbol{\theta} = \{(\rho_g, \pi_g, \delta_g) : g \in 1, \ldots, G\}$. *(b) Linking a series of snapshot HIV prevalence models at multiple time points* t*, to estimate HIV incidence in a ODE-driven compartmental model. Time* t *data* $\mathbf{y_t} = (y_t^1, y_t^2, \ldots, y_t^n)$ *are augmented by demographic and other data* $\mathbf{z_t} = (y_t^{n+1}, \ldots, y_t^N)$*, informing some of the transition rates* $\boldsymbol{\lambda}_t$*, such as migration and new HIV diagnoses. The parameters from (a), both basic and functional, are now encapsulated within* $\boldsymbol{\theta}_t$.

ciency, conflict assessment or uncertainty in model structure, it may be convenient to break the problem into smaller components, for example [28] fit a model for HCV prevalence in two stages. Although this 'modular' approach is often reasonable and computationally convenient, merging the resulting sub-models into a single analysis is non-trivial (see Section 4).

2.2 Dynamic Models

When interest is in estimating the temporal evolution of an epidemic, and the rates of infection in particular, dynamic models are necessary. There are two alternative approaches depending on the nature of the available information: linking the snapshot analyses of Section 2.1 over time; or using routine time series data on the sequelae of infection. In the first approach, at time t, the observational model is

$$\boldsymbol{Y}_t \sim p_{\boldsymbol{\theta}}\left(\cdot|\boldsymbol{X}_t\right),$$

and the snapshots are linked over time via some smoothing of the state variables $\boldsymbol{X}_t$. In the case of the HIV prevalence example (Figure 2), for a generic risk group g and a series of snapshots over time, this linkage is achieved by embedding a continuous-time multi-state model in the serial snapshot evidence synthesis [27]. The population is partitioned into disease states $\boldsymbol{X}_t$ and model dynamics are described by a system of ordinary differential equations. Time-varying transition rates, including HIV incidence, are the basic parameters $\boldsymbol{\phi} = \boldsymbol{\lambda}_t$, which are identifiable through the inclusion of additional demographic data $\boldsymbol{z}_t$, contributing to the likelihood as Poisson or binomial terms. The basic parameters $\boldsymbol{\theta}_t = (\rho_t, \pi_t, \delta_t)$ of the prevalence model are now deterministic functions $\boldsymbol{\theta}_t = f(\boldsymbol{X}_t)$ of the disease states in the dynamic model.

Such temporally-linked snapshot evidence syntheses can be used also to estimate state vectors, $\boldsymbol{X}_t$, that represent log-incidence, as in a study of toxoplasmosis [38], where temporal smoothing of the state vector is through a random walk, e.g

$$\log\left(\boldsymbol{X}_t\right) \sim \log\left(\boldsymbol{X}_{t-1}\right) + \boldsymbol{\phi}^T \boldsymbol{\epsilon}_{X,t}.$$

In the second (dynamic) approach, when the available data are time series

counts of clinical endpoints, back-calculation has been widely employed to estimate disease incidence by combining the time series with information on the time from infection to the end point (the incubation period). The basic convolution equation

$$\mu(t) = \int_0^t h(s) f(t-s) ds. \tag{2.1}$$

expresses the link between the rate of occurrence of a clinical end point, $\mu(t)$, the rate $h(\cdot)$ at which new infections occur and the distribution of the time from infection to the end point, $f(\cdot)$.

To estimate HIV incidence, equation (2.1), initially based on AIDS diagnoses, has been developed extensively to incorporate additional data, for example, to: improve identifiability of $h(\cdot)$ in the recent past [12]; identify recent infections amongst new diagnoses [e.g. 43]; and provide a more comprehensive description of the epidemic.

In particular, various discrete-time multi-state backcalculations have been proposed, where states are defined by CD4 cell counts [6, and references therein]. Through such an approach, estimation of the number of undiagnosed infections is possible, by incorporating data on HIV diagnoses and CD4 counts taken at diagnosis. In such models, the distribution $f(\cdot)$ in Equation (2.1) is characterised by progression rates through disease states and diagnosis probabilities, d_t. Together with incidence, h_t, these quantities are modelled by [6] using random walks and the backcalculation can be framed as a state-space model as in Equations (1.1) and (1.2) [40]. Here, the state vector, $\boldsymbol{X}_t = (h_t, d_t, \boldsymbol{E}_t)$, comprises the infection and diagnosis rates, as well as the state occupancies, $\boldsymbol{E}_t$. As new infections are assumed to occur according to a Poisson process, the likelihood is tractable when marginalised over the $\boldsymbol{E}_t$, which greatly improves the efficiency and accuracy with which inference on (h_t, d_t) can be drawn. In this case, the diagnoses are Poisson distributed and the CD4 data follow a multinomial distribution. The challenge here is to be able to incorporate additional sources of data, such as information from tests for recent infection performed on new diagnoses, whilst maintaining this tractability.

3. EVIDENCE SYNTHESIS IN MECHANISTIC TRANSMISSION MODELS

The classic approach to tracking the spread of an epidemic is through compartmental models that partition the population into Susceptible/Infected/Removed (SIR) states [3], or one of many similar variants. In the epidemic modelling literature these models are labelled as mechanistic transmission models. They differ from the multi-state models of Section 2 due to the explicit modelling of the transmission mechanisms, where rates of infection are a function of the prevalent number of infected and infectious individuals. The dynamics of such mechanistic models unfold according to a system of ordinary or stochastic differential equations or their discrete-time difference approximations.

3.1 Deterministic epidemic dynamics

Models with a deterministic state relationship, but for which states are imperfectly observed, can be expressed as:

$$\begin{aligned} \boldsymbol{X}_t &= f_{\boldsymbol{\phi}}\left(\boldsymbol{X}_{t-1}\right) \\ \boldsymbol{Y}_t &\sim p_{\boldsymbol{\theta}}\left(\cdot|\boldsymbol{X}_t\right), \end{aligned} \tag{3.1}$$

where $f_{\boldsymbol{\phi}}(\cdot)$ is a deterministic function, characterised by parameter $\boldsymbol{\phi}$, and $\boldsymbol{X}_t$ represents the distribution of the population in the SIR states, i.e. $\boldsymbol{X}_t = (S_t, I_t, R_t)$. Typically, $\boldsymbol{\phi}$ will include rates of transition between model states, relative rates of contact between different population strata and the transmission potential. Movements between model states will be unobserved, and, as in Section 2, the use of multiple data sources becomes necessary to identify both parameters and latent quantities. A number of examples exist where traditional epidemic surveillance information is augmented by additional serological, demographic, administrative or environmental data.

Surveillance and Serological Data Serological data, from testing of blood samples to detect the presence of antibodies, provide crucial information on the level of immunity in a population. The important role played by this type of data in uncovering an epidemic's dynamics is highlighted in applications to influenza data from Israel [42] and from England [8]. Due to the presence of asymptomatic infection, the magnitude of the epidemic cannot be estimated while the epidemic is ongoing from influenza-like illness data and associated virological swabbing alone. This idea is extended in [16], where changes in the immunity profile of a population and the fluctuating transmissibility of the virus between temporally distinct waves of infection are estimated.

In the language of transmission modelling, serological data $\boldsymbol{Y}_t^{\text{sero}}$ provide direct evidence on the number of people in the susceptible state. Incorporation of these data extends the observation model characterised by $p_{\boldsymbol{\theta}}$ in equation (3.1). The additional component, at time t, is typically binomial:

$$\boldsymbol{Y}_t^{\text{sero}}|\boldsymbol{X}_t \sim \text{Bin}\left(n_t^{\text{sero}}, p_t^{\text{sero}}\right),$$

where

$$p_t^{\text{sero}} = \mathbb{P}(\text{seropositive at time } t) = 1 - S_t/N$$

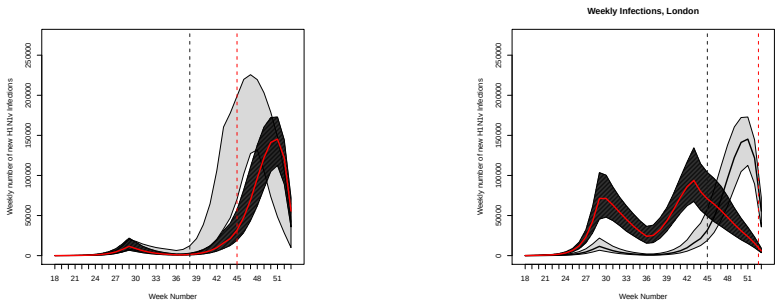
and n_t^{sero} is an assumed known sample size and N is the population size.

However, serological data can hold richer information than mere binary responses. In an application to the Dutch A/H1N1pdm influenza outbreak [37], data obtained from more sensitive micro-array assays are used to give a probabilistic interpretation of immunity. This is achieved via the specification of a mixture model for the log-titre values, classifying individuals into groups who are susceptible, recently infected or have long-held immunity. Here, the Y_t^{sero} are continuous responses distributed as

$$Y_t^{\text{sero}} \sim \frac{S_t}{N} p\left(\cdot|\boldsymbol{\theta}^{\text{s}}\right) + \frac{S_0 - S_t}{N} p\left(\cdot|\boldsymbol{\theta}^{\text{r}}\right) + \frac{N - S_0}{N} p\left(\cdot|\boldsymbol{\theta}^{\text{i}}\right),$$

where the $p(\cdot|\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ are normal density functions, corresponding to the distribution of log-titre values for susceptible (s), recently infected (r) and immune (i) sub-groups.

(a) No serological data, at $t = 178$ (b) No serological data, at $t = 245$



(c) Including serological data, at $t = 178$ (d) Including serological data, at $t = 245$

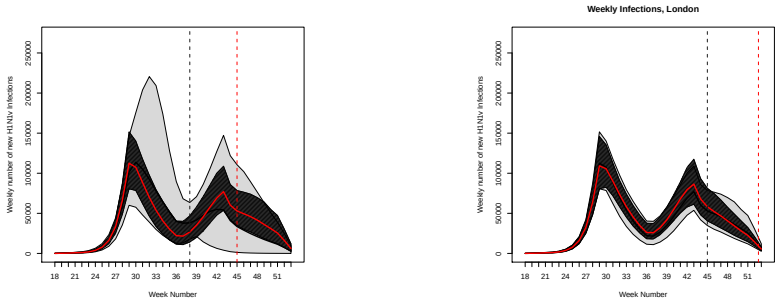


FIG 3. *Forecasts of the number of new A/H1N1pdm influenza infections after* $t = 178$ *and* 245 *days of 2009 pandemic data, in the absence ((a) and (b)) and presence ((c) and (d)) of serological data: posterior median (red central line); 95% credible interval (light grey region) for a forecast at a previous time (grey dashed vertical lines); 95% credible interval (dark grey region) for a 'current' forecast at t (red dashed vertical lines).*

The impact of serological data can be significant. Adapting figures from [8], Figures 3(a) and (b) show estimates and predictions of the number of new A/H1N1pdm influenza infections, when only data on syndromic consultations with a doctor are used. Analyses are carried out approximately three quarters of the way through and towards the end of the epidemic respectively, without any serological information. Figures 3(c) and (d) display the same results from analyses that additionally use the serological data. In the bottom row of Figure 3, epidemic projections appear to be nested as data accrue, with credible intervals narrowing. In the top row, a coherent picture of the epidemic is only obtained once the epidemic is almost over. In the absence of direct serological information on the number of infections, fitting a transmission model to doctor consultation data alone is of limited utility. A major epidemiological challenge, however, is to develop systems that can ensure the timely provision of these data during an ongoing pandemic.

Surveillance and Demographic, Administrative or Environmental Data An example of joint modelling of surveillance and demographic data is in [27], where the model in Figure 2(b) is extended to include a component of disease transmission utilising information $\boldsymbol{z}_t$ on ageing, migration and mortality. This is a rare example where such data are directly modelled, i.e. both $(\boldsymbol{Y}_t, \boldsymbol{Z}_t)$ have distributions, whereas more commonly, demographic data are treated as fixed covariates, rather than a joint outcome. In the latter case, the system equation in (3.1) is

replaced by

$$\boldsymbol{X}_t = f_{\boldsymbol{\phi}}\left(\boldsymbol{X}_{t-1}, \boldsymbol{Z}_t\right).$$

These explanatory data can come in many forms: [5] uses vaccination data to inform transition rates out of a susceptible state; [9] use commuting data to describe inter-region transmission; [42] relate transmission of A/H1N1pdm influenza in Israel to an index of 'mean absolute humidity'.

3.2 Stochastic epidemic dynamics

The full state-space specification of equations (1.1) and (1.2) is required in two contexts. The first context arises when the numbers of infected individuals are small enough for stochastic fluctuations in transmission to significantly impact on the future epidemic trajectory ('demographic stochasticity'). Statistical inference based on a model with deterministic dynamics can lead to poor forecasts for the timing of an epidemic peak and can preclude the possibility of epidemic extinction when $R > 1$, no matter how small the population of infected individuals. Secondly, deterministic dynamics are inadequate in the presence of environmental or other external factors not captured by the transmission model. Stochasticity in the temporal evolution of parameter values ('environmental stochasticity') can eliminate the possibility of over-optimistic, possibly biased forecasts that may otherwise result. Models that account for demographic stochasticity, such as, for example, the chain multinomial [40], model the evolution of the epidemic in discrete time. The evolution of the SIR-type disease states $\boldsymbol{X}_t$ forms a Markov process as in equation (1.1). However, the second context of environmental stochasticity is more prevalent in the literature. Here, mechanistic transmission models are driven by a time-varying transmission potential β_t, commonly modelled as a stochastic process. In [17] and [41], β_t is cast as Wiener and Gaussian processes respectively, whereas [34] impose a random effects model on the probability p_g of a susceptible individual in population subgroup g being infected within a chain-binomial model. The probability of a member of group g not being infected by any other infectious individual is expressed as

$$\left(1 - \frac{p_g}{N_g}\right)^{w_t \sum_{g'} C_{g',g} I_{g',t-1}},$$

where $w_t \sum_{g'} C_{g',g} I_{g',t-1}$ is the total number of infectious contacts experienced by a member of g, with C being a contact matrix and $I_{g',t-1}$ giving the time $t-1$ number of infectious individuals in strata g'. The correlated random effects, w_t, absorb any temporal fluctuations in infectivity and rates of contact. Here, due to the stratified population, the transmission potential has to be expressed for each type of contact, $\beta_t^{g,g'} = w_t(p_g C_{g',g}/N_g)$. A global value is derived as the dominant eigenvalue of a matrix $\boldsymbol{\beta}_t$, commonly known as the next-generation matrix, that has $\beta_t^{g,g'}$ as its $(g,g')^{\text{th}}$ entry.

The motivation for the use of multiple sources of data in stochastic epidemic modelling is no different to the deterministic case. However, there are fewer examples of their use.

Surveillance data Of these few examples, [34] constitutes a rare instance of using multiple epidemiological time series: the observations $\boldsymbol{y}_{1:T}$ comprise both laboratory-confirmed data on 'mild' cases and data on (nested) admissions to

hospitals and to ICUs. Both the types of stochasticity described above are incorporated. However, the complexity inherent in this model means that its run-time on a high-performance computing cluster is measured in months. Whilst this is not an impediment to retrospective epidemic analysis, it is deeply prohibitive for real-time analysis. Computational, potentially sequential, methods that enable a more swift use of such a model would be of great utility.

Surveillance and phylogenetic data The synthesis of genetic and epidemiological data is more common in the literature and is used to improve understanding of the transmission dynamics of a particular pathogen. Genetic sequence data (comprising the sequences themselves, together with associated sampling times) can allow reconstruction of transmission trees either by modelling the evolution of the pathogen explicitly using coalescent models to estimate the branching points of the trees [e.g 15, and references therein] or by using the genetic distance between the observed sequences [39]. The precise method depends on the assumptions that are appropriate for the pathogen and epidemic under investigation. These assumptions cover the possible presence of: within-host pathogen genetic variation; transmission bottlenecks (where a subset of the within-host variants are transmitted); unobserved cases; and introductions into the population. Attendant epidemiological data can add precision to the reconstruction of transmission trees, for example, by providing information on infectious periods or generation intervals, or on the dates at which particular individuals were at risk of infection [15].

There is an increasing body of work linking phylogenies into mechanistic transmission models. A general framework for identifying SIR and SEIR transmission models on the basis of phylogenetic data alone is developed in [24], additionally presenting an application incorporating time series data on removals from the population. Similarly, it is noted in [29] that phylogenetic information is of particular utility in the case where the surveillance data that are typically used to inform transmission modelling are highly noisy or only weakly informative. Their work demonstrates the improved estimation of epidemiological parameters possible when the analysis of epidemiological surveillance data using a continuous-time, continuous-space stochastic epidemic model is augmented by a sample of infection lineages.

As identified by [23] the challenge remains to relax many of the assumptions listed above for phylogenetic modelling, whilst incorporating additional aspects of outbreak dynamics. Consideration of an ever-increasing array of epidemiological data should make this a more achievable goal.

4. DISCUSSION

The recent increase in the number of evidence syntheses, mostly Bayesian, to estimate latent characteristics of epidemics is testimony of the crucial role of data from multiple sources. This role has been comprehensively explored in other reviews [1, 14], but briefly, include two key aims: identifiability of a wider range of (unobservable) quantities that can inform public health efforts to control epidemics than would be achievable from a single data source; and increased precision in estimates of these quantities, due to the use of all available relevant data, both direct and indirect. Advantages of Bayesian evidence synthesis include the ability to: introduce and formally quantify expert judgement in the form of

prior distributions; readily account for and estimate known biases in observational data through the introduction of bias parameters with carefully chosen priors; and minimise selection bias. However, the adoption of evidence synthesis methods, to achieve identifiability and precision, necessitates models of increasing realism and complexity, which are in turn accompanied by some general challenges that remain open questions [13], as we have highlighted through various examples in this review.

Complex models imply a need for various model building strategies, including hierarchical modelling for identifiability and modular approaches. How best to achieve identifiability from the currently available data is an active area of research. An algebraic determination, ahead of any inference, of parameter identifiability in a complex dynamic system has been explored recently in systems biology [e.g. 20]: such methods have the potential to be adapted to transmission modelling. A promising alternative is the extension of value-of-information methods to the evaluation of gains in precision in parameter estimates resulting from collecting or incorporating further evidence, proposed in application to the HIV prevalence context in [22].

Reasons for a modular approach, dividing a complex model into smaller sub-models, include: understanding the influence of each evidence source on joint inference; assessing and resolving conflict during the model building process; and computational tractability. However, incorporating the results of each sub-model into a second-stage joint model in a manner that retains the feedback from different data sources to common parameters is not straightforward. Recent work that allows for principled inference from a fully joint model given posterior samples from sub-models has been proposed [19]. The application of this "Markov melding" approach to evidence syntheses has the potential to facilitate the increasingly realistic and complex models required in the stochastic epidemic field.

The potential for conflicting evidence is a challenge, but evidence synthesis provides a framework in which, once any conflict has been detected, measured and resolved, models are internally validated: an adequate final model is consistent with every data source included. However, systematic cross-validatory conflict assessment [13] as with any modular approach, is computationally intensive: adaptation is needed to enable timely inference. Conflict resolution through, for example, bias modelling and evidence weighting methods, is a next step [13]. However, while in a frequentist framework there are well-established methods to account for selection biases in the types of observational data usually included in epidemic evidence syntheses, Bayesian equivalents are still in their infancy [36].

A recurring theme through each of the above challenges is that of computationally efficient statistical inference. In the context of epidemic modelling, timely estimation is crucial to address public health policy needs in the midst of an emerging epidemic [13]. Much progress has been made in developing and applying efficient algorithms for epidemic evidence syntheses, such as: sequential Bayesian methods [33, 7], including likelihood-free particle MCMC [29]; and approximate Bayesian computation [30]. Alternatively, to achieve computational efficiency, one might approximate the complex epidemic model with a readily implementable proxy. Shaman and colleagues have extensively used an extended Kalman filter [e.g. 32], to provide a stochastic time series approximation to the dynamics of SIR models. Another approach is Bayesian emulation [18], which

seeks to characterise an epidemic model with an emulator, built from a dynamic Gaussian process prior. A similar emulation approach is adopted by [4], who use history matching to calibrate a complex, multi-output epidemic simulation model. This latter work is an attempt to tackle the next challenge, to broaden the scope of all such algorithms to handle multiple datasets, possibly diverse in nature.

4.1 Conclusions

A recent review of infectious disease modelling [26] suggests that the full potential of mechanistic models that "simultaneously link data from diverse, heterogeneous data sources" has yet to be reached. This is certainly true for fully stochastic transmission models, though rare examples of such models embedded within an evidence synthesis do exist [30, 34]. Such rarity and the challenges discussed above motivate the need for further development in this area.

However, the many examples reviewed in Section 3.1, particularly for deterministic models, suggest that evidence synthesis for mechanistic models is both a well-established and rapidly expanding field.

REFERENCES

[1] Ades, A. E. and Sutton, A. J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *J R Stat Soc Ser A Stat Soc* **169** 5–35.

[2] Albert, I., Espié, E., De Valk, H. and Denis, J. B. (2011). A Bayesian Evidence Synthesis for Estimating Campylobacteriosis Prevalence. *Risk Anal.* **31** 1141–1155.

[3] Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

[4] Andrianakis, I., McCreesh, N., Vernon, I., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M. and White, R. G. (2017). Efficient History Matching of a High Dimensional Individual-Based HIV Transmission Model. *SIAM/ASA J. Uncertain. Quantif.* **5** 694–719.

[5] Baguelin, M., Flasche, S., Camacho, A., Demiris, N., Miller, E. and Edmunds, W. J. (2013). Assessing Optimal Target Populations for Influenza Vaccination Programmes: An Evidence Synthesis and Modelling Study. *PLoS Medicine* **10** e1001527+.

[6] Birrell, P. J., Chadborn, T. R., Gill, O. N., Delpech, V. C. and De Angelis, D. (2012). Estimating Trends in Incidence, Time-to-Diagnosis and Undiagnosed Prevalence using a CD4-based Bayesian Back-calculation. *Stat. Commun. Infec. Dis.* **4**.

[7] Birrell, P. J., De Angelis, D., Wernisch, L., Tom, B. D. M., Roberts, G. O. and Pebody, R. G. (2016). Efficient real-time monitoring of an emerging influenza epidemic: how feasible? *arXiv preprint* http://arxiv.org/abs/1608.05292.

[8] Birrell, P. J., Ketsetzis, G., Gay, N. J., Cooper, B. S., Presanis, A. M., Harris, R. J., Charlett, A., Zhang, X.-S., White, P. J., Pebody, R. G. and De Angelis, D. (2011). Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proc. Natl. Acad. Sci. USA* **108** 18238–18243.

[9] Birrell, P. J., Zhang, X.-S., Pebody, R. G., Gay, N. J. and De Angelis, D. (2016). Reconstructing a spatially heterogeneous epidemic: Characterising the geographic spread of 2009 A/H1N1pdm infection in England. *Sci. Rep.* **6** 29004.

[10] Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer.

[11] Conti, S., Presanis, A. M., van Veen, M. G., Xiridou, M., Donoghoe, M. C., Stengaard, A. R. and De Angelis, D. (2011). Modeling of the HIV infection epidemic in the Netherlands: A multi-parameter evidence synthesis approach. *Ann. Appl. Stat.* **5** 2359–2384.

[12] De Angelis, D. (2011). Back-calculation. In *Encyclopaedic companion to medical statistics* 2 ed. (B. S. Everitt and C. Palmer, eds.) 23. John Wiley & Sons.

[13] De Angelis, D., Presanis, A. M., Birrell, P. J., Scalia Tomba, G. and House, T. (2014). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* **10** 83–87.

[14] De Angelis, D., Presanis, A. M., Conti, S. and Ades, A. E. (2014). Estimation of HIV burden through Bayesian evidence synthesis. *Statist. Sci.* **29** 9–17.

[15] De Maio, N., Wu, C.-H. and Wilson, D. J. (2016). SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLOS Comput. Biol.* **12** e1005130.

[16] Dorigatti, I., Cauchemez, S. and Ferguson, N. M. (2013). Increased transmissibility explains the third wave of infection by the 2009 H1N1 pandemic virus in England. *Proc. Natl. Acad. Sci. USA* **110** 13422–13427.

[17] Dureau, J., Kalogeropoulos, K. and Baguelin, M. (2013). Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics* **14** 541–555.

[18] Farah, M., Birrell, P., Conti, S. and Angelis, D. D. (2014). Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza. *J. Amer. Statist. Assoc.* **109** 1398–1411.

[19] Goudie, R. J. B., Presanis, A. M., Lunn, D., De Angelis, D. and Wernisch, L. (2016). Model surgery: joining and splitting models with Markov melding. *arXiv preprint* http://arxiv.org/abs/1607.06779.

[20] Gross, E., Harrington, H. A., Rosen, Z. and Sturmfels, B. (2016). Algebraic Systems Biology: A Case Study for the Wnt Pathway. *Bull. Math. Biol.* **78** 21–51.

[21] Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K. T. D., Edmunds, W. J., Frost, S. D. W., Funk, S., Hollingsworth, T. D., House, T., Isham, V., Klepac, P., Lessler, J., Lloyd-Smith, J. O., Metcalf, C. J. E., Mollison, D., Pellis, L., Pulliam, J. R. C., Roberts, M. G. and Viboud, C. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science* **347** aaa4339-aaa4339.

[22] Jackson, C., Presanis, A., Conti, S. and De Angelis, D. (2017). Value of Information: Sensitivity Analysis and Research Design in Bayesian Evidence Synthesis. *arXiv preprint arXiv:1703.08994*.

[23] Klinkenberg, D., Backer, J. A., Didelot, X., Colijn, C. and Wallinga, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Comput. Biol.* **13** e1005495.

[24] Lau, M. S. Y., Marion, G., Streftaris, G., Gibson, G., Chase-Topping, M. and Haydon, D. (2015). A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLOS Comput. Biol.* **11** e1004633.

[25] Lauritzen, S. L. (1996). *Graphical models* **17**. Clarendon Press.

[26] Lessler, J., Azman, A. S., Grabowski, M. K., Salje, H. and Rodriguez-Barraquer, I. (2016). Trends in the Mechanistic and Dynamic Modeling of Infectious Diseases. *Curr Epidemiol Rep.* 212–222.

[27] Presanis, A. M., De Angelis, D., Goubar, A., Gill, O. N. and Ades, A. E. (2011). Bayesian evidence synthesis for a transmission dynamic model for HIV among men who have sex with men. *Biostatistics* **12** 666–681.

[28] Prevost, T. C., Presanis, A. M., Taylor, A., Goldberg, D. J., Hutchinson, S. J. and De Angelis, D. (2015). Estimating the number of people with hepatitis C virus who have ever injected drugs and have yet to be diagnosed: An evidence synthesis approach for Scotland. *Addiction* **110** 1287–1300.

[29] Rasmussen, D. A., Ratmann, O. and Koelle, K. (2011). Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. *PLoS Comput Biol* **7** e1002136+.

[30] Ratmann, O., Donker, G., Meijer, A., Fraser, C. and Koelle, K. (2012). Phylodynamic Inference and Model Assessment with Approximate Bayesian Computation: Influenza as a Case Study. *PLoS Comput Biol* **8** e1002835+.

[31] Rosinska, M., Gwiazda, P., De Angelis, D. and Presanis, A. M. (2016). Bayesian evidence synthesis to estimate HIV prevalence in men who have sex with men in Poland at the end of 2009. *Epidemiol. Infect.* **144** 1175–1191.

[32] Shaman, J., Karspeck, A., Yang, W., Tamerius, J. and Lipsitch, M. (2013). Real-time influenza forecasts during the 2012–2013 season. *Nat Commun.* **4** 2837 EP -.

[33] Sheinson, D. M., Niemi, J. and Meiring, W. (2014). Comparison of the performance of particle filter algorithms applied to tracking of a disease epidemic. *Math. Biosci.* **255**

21–32.

[34] Shubin, M., Lebedev, A., Lyytikäinen, O. and Auranen, K. (2016). Revealing the True Incidence of Pandemic A(H1N1)pdm09 Influenza in Finland during the First Two Seasons - An Analysis Based on a Dynamic Transmission Model. *PLoS Comp. Biol.* **12** e1004803.

[35] Shubin, M., Virtanen, M., Toikkanen, S., Lyytikäinen, O. and Auranen, K. (2014). Estimating the burden of A(H1N1)pdm09 influenza in Finland during two seasons. *Epidemiol. Infect.* **142** 964–974.

[36] Si, Y., Pillai, N. S. and Gelman, A. (2015). Bayesian Nonparametric Weighted Sampling Inference. *Bayesian Anal.* **10** 605–625.

[37] te Beest, D. E., Birrell, P. J., Wallinga, J., De Angelis, D. and van Boven, M. (2015). Joint modelling of serological and hospitalization data reveals that high levels of pre-existing immunity and school holidays shaped the influenza A pandemic of 2009 in the Netherlands. *J Royal Soc Interface* **12** 20141244-.

[38] Welton, N. J. and Ades, A. E. (2005). A model of toxoplasmosis incidence in the UK: Evidence synthesis and consistency of evidence. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 385–404.

[39] Worby, C. J., O'Neill, P. D., Kypraios, T., Robotham, J. V., De Angelis, D., Cartwright, E. J. P., Peacock, S. J. and Cooper, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann. Appl. Stat.* **10** 395–417.

[40] Wu, H. and Tan, W. Y. (2000). Modelling the HIV epidemic: A state-space approach. *Math. Comput. Model.* **32** 197–215.

[41] Xu, X., Kypraios, T. and O'Neill, P. D. (2016). Bayesian non-parametric inference for stochastic epidemic models using Gaussian Processes. *Biostatistics* **17** 619–633.

[42] Yaari, R., Katriel, G., Stone, L., Mendelson, E., Mandelboim, M. and Huppert, A. (2016). Model-based reconstruction of an epidemic using multiple datasets: understanding influenza A/H1N1 pandemic dynamics in Israel. *J Royal Soc Interface* **13** 92–92.

[43] Yan, P., Zhang, F. and Wand, H. (2011). Using HIV Diagnostic Data to Estimate HIV Incidence: Method and Simulation. *Stat. Commun. Infec. Dis.* **3** -.