Epidemics 10 (2015) 83–87

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Four key challenges in infectious disease modelling using data from multiple sources

Daniela De Angelis[a,b,*], Anne M. Presanis[a], Paul J. Birrell[a], Gianpaolo Scalia Tomba[c], Thomas House[d]

[a] MRC Biostatistics Unit, Cambridge Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK
[b] Public Health England, 61 Colindale Avenue, London NW9 5HT, UK
[c] Department of Mathematics, University of Rome Tor Vergata, Rome, Italy
[d] Warwick Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK

ARTICLE INFO

Article history:
Received 21 February 2014
Received in revised form 6 August 2014
Accepted 16 September 2014
Available online 28 September 2014

Keywords:
Evidence synthesis
Bayesian
Statistical inference
Multiple sources
Epidemics
Complex models

ABSTRACT

Public health-related decision-making on policies aimed at controlling epidemics is increasingly evidence-based, exploiting multiple sources of data. Policy makers rely on complex models that are required to be robust, realistically approximating epidemics and consistent with all relevant data. Meeting these requirements in a statistically rigorous and defendable manner poses a number of challenging problems. How to weight evidence from different datasets and handle dependence between them, efficiently estimate and critically assess complex models are key challenges that we expound in this paper, using examples from influenza modelling.

Introduction

Increasingly, there is a perceived need to exploit information from multiple sources in epidemic modelling, ensuring decision-making on public health policies geared to control epidemics is progressively based on as many diverse sources of information as possible (Rutherford et al., 2010) and the use of models (e.g. https://www.gov.uk/government/policy-advisory-groups/joint-committee-on-vaccination-and-immunisation). Policy makers want 'defendable' models that not only realistically approximate the phenomenon of interest, but are also, crucially, able to produce outputs consistent with all relevant available data (Rolka et al., 2007; Lipsitch et al., 2011). This requirement, supported by the continued progress in computational power, has encouraged the development of increasingly complex models, which, in turn, require rich arrays of data to guarantee parameter identifiability (Ferguson et al., 2006).

In addition, irrespective of the complexity of the model, modellers are often faced with the task of integrating information from many heterogeneous sources of data. For example, the behaviour of an epidemic in its early stages is described by the parameter R_0, the basic reproductive number. However, equally crucial for the containment of an infectious disease outbreak (Fraser et al., 2004; Powers et al., 2011) is knowledge of the proportion of transmission occurring before the onset of symptoms, θ. Population incidence data contain information on R_0, but are uninformative about θ. Complementary evidence from 'challenge' studies, where the time between infection and symptom onset is measured directly and information is available on the distributions of latent and infectious periods, are needed to estimate θ. A comprehensive description of the evolution of an outbreak can only be obtained using data from multiple sources.

It is, however, not typically the case that there will be a single data source directly informing each relevant parameter. More realistically, there will be a collection of datasets, each of different quality, that will need to be appropriately synthesised to derive the estimates of interest, as illustrated in Fig. 1. Here the epidemic process is modelled in terms of the basic parameters of interest, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$ and the information from each data source x_j, $j = 1, \ldots, n$, is expressed as a function of the basic parameters i.e. $\boldsymbol{\theta}_j^* = f_j(\boldsymbol{\theta})$. The form of this function, whether deterministic or stochastic, defines the relationship of the observation model to the epidemic model. Examples of $f_j(\boldsymbol{\theta})$ include cases where a data source provides: direct information on a single parameter of interest (i.e. $\theta_j^* = \theta_i$); biased evidence on $\boldsymbol{\theta}$ (see Section "Model criticism"); simultaneous

* Corresponding author at: MRC Biostatistics Unit, Cambridge Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK. Tel.: +44 1223330390.
E-mail address: daniela.deangelis@mrc-bsu.cam.ac.uk (D. De Angelis).

http://dx.doi.org/10.1016/j.epidem.2014.09.004

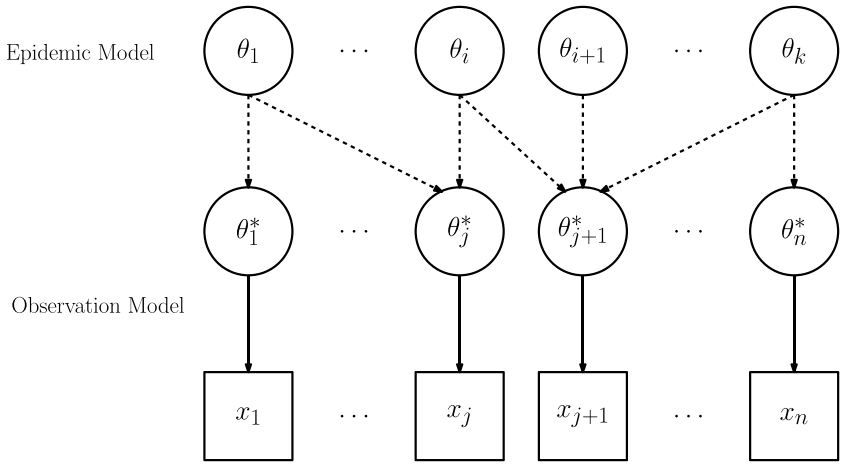

Fig. 1. Schematic diagram of how multiple data sources can link into an epidemic model via an observation model(s).

information on multiple components of $\boldsymbol{\theta}$ or on further nuisance parameters $\boldsymbol{\phi}$ (i.e. $\boldsymbol{\theta}_j^* = f_j(\boldsymbol{\theta}, \boldsymbol{\phi})$).

Estimation involves a flow backwards from the combined information to $\boldsymbol{\theta}$. Carrying out such inference in a principled manner is not straightforward and poses a number of challenges stemming from the multiplicity and the limitations in the available data sources. We illustrate the main ones below using mainly examples from recent literature on influenza, pointing out relevant ideas from the statistical literature that could be explored to address these challenges. Although, in principle, this type of synthesis can be carried out via maximum likelihood methods (e.g. Commenges and Hejblum, 2013), we mainly concentrate on a Bayesian approach as it represents a very natural approach to data assimilation both from a principled and computational point of view.

1. How should evidence be weighted?

When a multiplicity of data is used, the various sources of evidence will inevitably be of different quality and a natural question is whether and how to account for this diversity in the model (Ypma et al., 2012). Clearly the first challenge is to define 'quality'. Here 'quality' relates to both measurement error and bias. One immediate solution to the heterogeneity of quality would be to exclude the lower quality data with, however, a resulting loss of information and risk of introducing biases due to the selective nature of information retained (Turner et al., 2009). Alternatively, a few ways of weighting data can be explored, each posing its own challenges.

The most natural approach is through an appropriate choice of distributional assumption for each data item. For example, when analysing count data, contrast the use of a negative binomial likelihood with the Poisson, as was employed in two of the transmission models developed to estimate the evolution of the 2009 A/H1N1 influenza pandemic (Birrell et al., 2011; Dorigatti et al., 2012). Dorigatti et al. (2012), in particular, demonstrate the sensitivity of estimates of R_0 to the assumption of over-dispersion in the data. Furthermore, even within a specific distributional form, the degree to which error variance is modelled can have an impact upon the relative importance of each data component. This aspect of weighting of information is very closely linked to Section "Model criticism", as the correct assumption can be examined through methods for model choice.

A further approach is to recognise and model explicitly the limitations in the data, in particular in relation to bias (e.g. see recent criticism of Google 'Flu Trends by Olson et al., 2013). The observational model can be expanded to include additional parameters formally expressing such limitations. Magnitude and direction of the likely bias are incorporated through a suitable choice of a prior distribution for a bias parameter (Turner et al., 2009). This distribution ideally should be informative, at least in terms of the direction of the bias, to prevent the new parameter from absorbing all the unexplained variability, without offering any specific explanation for the nature of the bias. However, much remains to be done in terms of bias modelling, in particular in relation to self-reported data or data collected through particular channels, such as the Internet.

The concept of power priors (Chen and Ibrahim, 2000) represents an additional interesting avenue to be explored in the problem of weighting evidence. The principle comes from the world of clinical trials and has been proposed as an approach to incorporate data from a previous trial as an input to the analysis of a current study. The same concept could be applied to concurrent data sources, and the choice of appropriate values for the weighting scheme would be driven by expert opinion on the validity of each source or, perhaps, estimated, although this is still controversial (Neuenschwander et al., 2009).

General recommendations for the best strategy for the weighting of information do not exist, but formal thinking on how to approach such weighting of data should be encouraged as it is a choice to which modelling outcomes are rarely robust.

2. Handling dependence between datasets

In most cases where a multiplicity of datasets are used to inform a model, there will be some degree of dependency between them. Given a model, the important distinction is between datasets that are conditionally independent and those that are conditionally dependent. In the directed acyclic graph (Lauritzen, 1996) in Fig. 1, the datasets x_j, $j = 1, \ldots, n$ are independent, conditional on the model parameters $\boldsymbol{\theta}$, where the independence is represented by the lack of links between the x_js. This conditional independence is a common model assumption in many examples (e.g. Rasmussen et al., 2011; Strelioff et al., 2013). However, there might be situations in which the independence assumption is not tenable. An example of such data can be found in the surveillance of the 2009 influenza pandemic in the UK. Two transmission models (Birrell et al., 2011; Dorigatti et al., 2013) used, amongst other data sources, data on individuals consulting general practitioners (GPs) for influenza-like-illness (ILI). An additional relevant data source was the National Pandemic 'Flu Service (NPFS) (Evans et al., 2011), an internet and telephone service for the recording of self-reported symptoms and anti-viral distribution. It is possible that