Statistical challenges in epidemic modelling

Problemi inferenziali nella stima di modelli per epidemie

Daniela De Angelis, Paul Birrell and Anne Presanis

Abstract Health-related policy decision-making for epidemic control is increasingly evidence-based, exploiting multiple sources of data. Policy makers rely on models, which are required to be approximating realistically the process of interest and use all relevant information. This requirement poses a number of statically challenging problems. We review some of these challenges in this paper.

Abstract *Le strategie per il controllo di un'epidemia sono sempre piu' fondate su risultati di modelli matematici adottati per approssimare l'evoluzione dell' epidemia. Tali modelli sono pertanto attentamente scrutinati per il loro realismo e la capacita' di produrre risultati consistenti con tutte le informazioni disponibili. Da un punto vista inferenziale, l'uso di dati da multiple fonti pone una serie di problemi metodologici. In questo lavoro esponiamo alcuni di tali problemi.*

Key words: Inference, Multiple Sources, Evidence synthesis

In epidemic modelling, it is typically assumed that a latent system can approximate the epidemic process of interest. The dynamics of such a system are governed by

Daniela De Angelis
Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Robinson Way, Cambridge CB2 0SR, and Public Health England, 61 Colindale Avenue, London, NW9 5HT, UKe-mail: daniela.deangelis@mrc-bsu.cam.ac.uk

Paul Birrell
Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK e-mail: Paul.Birrell@mrc-bsu.cam.ac.uk

Anne Presanis
Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK e-mail: Anne.Presanis@mrc-bsu.cam.ac.uk

parameters representing quantities with a relevant interpretation. Typically, in the infectious disease literature, the study of the characteristics of the system and its future evolution has been carried out by simulation. The parameters are fixed to specific values, generally derived from previous studies or from assumptions, and the behaviour of the system is simulated. If any of the parameters is considered uncertain, it is assumed that they follow a suitable distribution from which values are sampled at each simulation run. This adds an appropriate degree of uncertainty to the results of the study. In the last decade, epidemic models have become increasingly central to health related decision making, including vaccination strategies or policies geared to control of epidemics. Therefore, the methodology underpinning epidemic modelling has been increasingly scrutinized. This, together with the development of suitable inferential techniques, has progressively led to the adoption of rigorous approaches to linking models to data [5].

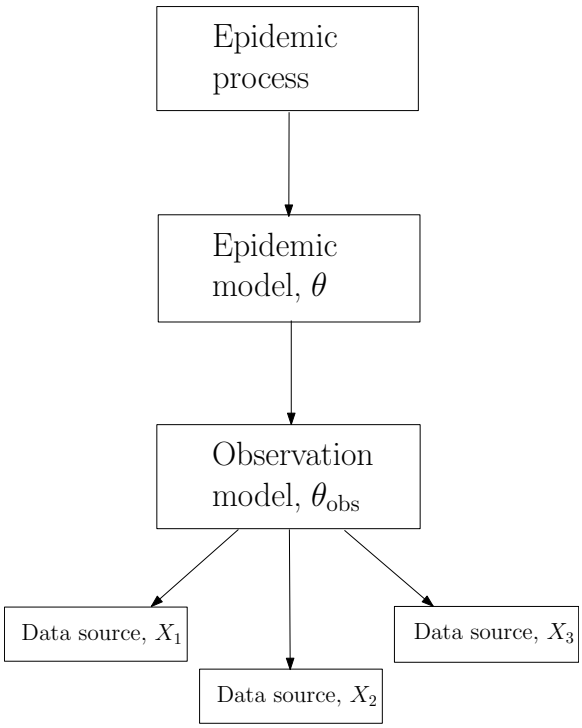

Fig. 1 Schematic diagram of how multiple data sources can link into an epidemic model via an observation model(s). The models are respectively parameterised by θ and θ_{obs}.

Continued progress in computational power has also encouraged the development of increasingly complex models, requiring rich arrays of data to guarantee parameter identifiability. This revolution has been encouraged by the pressing need, perceived by policy makers, to have 'defendable' models, *i.e.* able to approximate

realistically the process of interest and, crucially, be able to be consistent with all available information [4].

As a consequence, there has been a proliferation in the literature on infectious diseases of models incorporating data from multiple sources (see *e.g.* [6],[7], [8], [9]) for recent examples in the area of influenza) . From an inferential point of view, use of multiple sources of information addresses identifiability problems; minimises the risk of bias due to arbitrary exclusion of potentially relevant data; allows a more realistic assessment of uncertainty; and typically leads to more accurate estimation [1].

Figure 1 schematically represent how multiple data sources can be related to an epidemic process of interest. The epidemic process is modelled in terms of the basic parameters of interest, θ, however, the information from different data source X_i, $i = 1,2,3$, is typically only indirectly related to the basic parameters θ, by informing a function $\theta_i^* = f_i(\theta)$ of the basic parameters. Examples of $f_i(\theta)$ include the case when a data source provides biased information on θ or informs simultaneously multiple components of the parameter θ. The estimation problem involves the combination of the information on θ contained in each source.

The assimilation of information from multiple sources in a set up like the one in Figure 1 is not straightforward and poses a number of challenges stemming from the multiplicity of information. These challenges include the critical appraisal of the available information with the likely need to weight evidence according to its quality; the handling of potential dependency between data sources; the potential for lack of consistency in the inference from the different sources; and the computationally efficient inference of the resulting complex models [3]. Here we use the example of recent work on the H1N1 pandemic in England [2] to illustrate some of these challenges, which are exacerbated when inference on the evolution of an epidemic is required in real time, as data from multiple sources accumulate.

References

1. Ades, A. E., Sutton, A.J. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches Journal of the Royal Statistical Society: Series A (Statistics in Society) **169**, 3-5, (2006)
2. Birrell, Paul J. and Ketsetzis, Georgios and Gay, Nigel J. and Cooper, Ben S. and Presanis, Anne M. and Harris, Ross J. and Charlett, André and Zhang, Xu-Sheng and White, Peter J. and Pebody, Richard G. and De Angelis, Daniela.: Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. Proceedings of the National Academy of Sciences.**45**, 18238–18243, (2011)
3. De Angelis, Daniela, Presanis, Anne M., Birrell, Paul J., Scalia Tomba, Gianpaolo, House, Thomas. Five challenges in infectious disease modelling using data from multiple sources. Submitted.
4. Rutherford, George and McFarland, William and Spindler, Hilary and White, Karen and Patel, Sadhna and Grasse, John A. and Sabin, Keith and Smith, Nathan and Tache, Stephanie and Garcia, Jesus C. and Stoneburner, Rand. : Public health triangulation: approach and application to synthesizing data to understand national and local HIV epidemics. BMC Public Health. **10**, 447, (2010)

5. O'Neill Philip D. Introduction and snapshot review: Relating infectious disease transmission models to data. Statistics in Medicine **29**, 2069–2077, (2010)
6. Dorigatti, Ilaria and Cauchemez, Simon and Ferguson, Neil M. Increased transmissibility explains the third wave of infection by the 2009 H1N1 pandemic virus in England. Proceedings of the National Academy of Sciences **110**, 13422–13427, (2013)
7. Rasmussen, David A. and Ratmann, Oliver and Koelle, Katia. Inference for Nonlinear Epidemiological Models Using Genealogies and Time Serie. PLoS Comput Biology. (2011) doi: 10.1371/journal.pcbi.100213.
8. Strelioff, Christopher C. and Vijaykrishna, Dhanasekaran and Riley, Steven and Guan, Yi and Peiris, J. S. Malik and Lloyd-Smith, James O. Inferring patterns of influenza transmission in swine from multiple streams of surveillance data. Proceedings of the Royal Society B: Biological Sciences **280**, (2013)
9. Ypma, R. J. F. and Bataille, A. M. A. and Stegeman, A. and Koch, G. and Wallinga, J. and van Ballegooijen, W. M. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proceedings of the Royal Society B: Biological Science **279**, 444–450, (2012)