Preventive Veterinary Medicine 122 (2015) 213–220

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Spatial and temporal epidemiological analysis in the Big Data era

Dirk U. Pfeiffer*, Kim B. Stevens

Veterinary Epidemiology, Economics & Public Health Group, Department of Production & Population Health, Royal Veterinary College, London, UK

ARTICLE INFO

Article history:
Received 3 February 2015
Received in revised form 27 May 2015
Accepted 31 May 2015

Keywords:
Data science
Exploratory analysis
Internet of Things
Modelling
Multi-criteria decision analysis
Spatial analysis
Visualisation

ABSTRACT

Concurrent with global economic development in the last 50 years, the opportunities for the spread of existing diseases and emergence of new infectious pathogens, have increased substantially. The activities associated with the enormously intensified global connectivity have resulted in large amounts of data being generated, which in turn provides opportunities for generating knowledge that will allow more effective management of animal and human health risks. This so-called *Big Data* has, more recently, been accompanied by the *Internet of Things* which highlights the increasing presence of a wide range of sensors, interconnected via the Internet. Analysis of this data needs to exploit its complexity, accommodate variation in data quality and should take advantage of its spatial and temporal dimensions, where available. Apart from the development of hardware technologies and networking/communication infrastructure, it is necessary to develop appropriate data management tools that make this data accessible for analysis. This includes relational databases, geographical information systems and most recently, cloud-based data storage such as Hadoop distributed file systems. While the development in analytical methodologies has not quite caught up with the *data deluge*, important advances have been made in a number of areas, including spatial and temporal data analysis where the spectrum of analytical methods ranges from visualisation and exploratory analysis, to modelling. While there used to be a primary focus on statistical science in terms of methodological development for data analysis, the newly emerged discipline of *data science* is a reflection of the challenges presented by the need to integrate diverse data sources and exploit them using novel data- and knowledge-driven modelling methods while simultaneously recognising the value of quantitative as well as qualitative analytical approaches. Machine learning regression methods, which are more robust and can handle large datasets faster than classical regression approaches, are now also used to analyse spatial and spatio-temporal data. Multi-criteria decision analysis methods have gained greater acceptance, due in part, to the need to increasingly combine data from diverse sources including published scientific information and expert opinion in an attempt to fill important knowledge gaps. The opportunities for more effective prevention, detection and control of animal health threats arising from these developments are immense, but not without risks given the different types, and much higher frequency, of biases associated with these data.

1. Introduction

Economic and technological developments in the last 50 years have led to global eco-social system changes that greatly facilitate the emergence and spread of infectious diseases in both animals and humans. This represents a major challenge for the management of infectious disease risks and is likely to require a paradigm shift in analytical approaches rather than an evolution of existing ones. This change in approach is reflected in the widespread recognition of the need to adopt inter- and transdisciplinary approaches in risk research and management. In addition, the digital revolution has provided major opportunities with respect to data collection and analysis. This has now evolved into the Internet of Things where everyday objects are connected through information networks, allowing them to send and receive data (Anon., 2014b; Kamel Boulos and Al-Shorbaji, 2014). Related to this, is the so-called *Industry 4.0* (a collective term for technologies and concepts of value chain organisation; (Lee et al., 2014)), which reflects a vision for how the industrial sector may respond to the tight integration between the physical and digital world through the implementation of smart value chains.

* Corresponding author at: Veterinary Epidemiology, Economics & Public Health Group, Dept. of Production & Population Health, Royal Veterinary College, Hawkshead Lane, Hatfield, Hertfordshire, AL97TA, United Kingdom. Fax: +44 1707 666574.
E-mail address: pfeiffer@rvc.ac.uk (D.U. Pfeiffer).

http://dx.doi.org/10.1016/j.prevetmed.2015.05.012
0167-5877/

The concepts of smart health (Solanas et al., 2014), mHealth (Istepanian et al., 2004) and eHealth (Eysenbach, 2001) can be seen as the starting point for these developments and, together with the recent increase in popularity, and availability, of wearable sensors, have boosted the development of associated technologies. However, these sensors, other measurement devices and data sources are of limited use if the raw data they generate are not converted into information that can inform decision making, which has led to the need for suitable data management and analytical methods that can handle the resulting large, heterogeneous datasets.

In animal health in general, and veterinary epidemiology specifically, the established methodological frameworks provide guidance for research of cause-effect relationships based on data generated through *a priori* designed field and laboratory studies. This review explores recent developments, and future directions, for spatial and temporal analysis in support of managing complex animal health problems. We begin this review from a broader perspective by focussing on the developments that have led to the data revolution and its impact on the health sciences. We then discuss how the new scientific discipline of data science has been established to tackle the analytical challenges and opportunities resulting from the data revolution. From this wider analytical context, we then focus on the specific developments in spatio-temporal epidemiological data analysis resulting from the data revolution.

2. Data revolution: from the Internet via Big Data to the Internet of Things

Scientific approaches aimed at improving our understanding of the complexity of the systems of which animal and human diseases form a part, usually involve data collection. However, the way in which data are generated has changed radically over the last 30 years, mainly as a result of the emergence of electronic methods for measuring, recording, storing and distributing data. As part of this development, the Internet now forms the backbone of a globally-reaching information network. The drivers behind the data revolution have been multiple, and early on were dominated by defence, public safety and scientific interests. Only once commercial companies such as Google (https://www.google.com), Amazon (http://www.amazon.com) and Facebook (https://www.facebook.com) were able to demonstrate, during the last 10 years, the potential for commercial exploitation, did the data revolution truly take off. There are also now increasing concerns in relation to potential abuse of Big Data (Schadt, 2012; Anon., 2014a).

Mayer-Schönberger and Cukier (2014) define Big Data as '*The ability of society to harness information in novel ways to produce useful insights or goods and services of significant value*' and '*. . .things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value*.' Big data are generally characterised by 3Vs: volume (relative magnitude of dataset), velocity (rate at which new data are generated) and variety (heterogeneous structure of dataset [e.g. text, video, audio]) (Gandomi and Haider, 2015). A fourth 'v' frequently used to describe Big Data is veracity which acknowledges the inherent uncertainty frequently associated with, in particular, web-based Big Data and the corresponding need for analytical approaches that are able to account for this unreliability (Gandomi and Haider, 2015). In addition, the business community has added a fifth 'v'; *value*.

Traditional database management systems based on tabular or relational data management structures are not suited to dealing with Big Data as most of it is unstructured. Cloud-based data storage using the Apache Hadoop® distributed file system (http://hadoop.apache.org; last accessed April 2015) has been developed to allow efficient management of such data (O'Driscoll et al., 2013; Fernández et al., 2014). A data mining approach was used to explore the use of search term data for prediction of flu trends (Ginsberg et al., 2009) based on the assumption that changes in information and communication patterns on the Internet can act as early warning of changes in population health (Wilson and Brownstein, 2009). This resulted in the development of the search-term surveillance system, Google Flu Trends (GFT; http://www.google.org/flutrend, last accessed April 2015). By combining data-mining of Google search queries and statistical modelling, GFT provides a baseline indicator of the trend or changes in the rate of influenza, thereby providing estimates of weekly regional US influenza activity with a reporting lag of only one day compared with the 1–2 week delays associated with the Centers for Disease Control and Prevention (CDC) Influenza Sentinel Provider Surveillance reports (Ginsberg et al., 2009). However, the results generated by this algorithm have been the subject of controversy as predictions were incorrect at specific time points when they particularly mattered (Butler, 2013; Lazer et al., 2014). The fact remains though, that the relative immediacy of web-based surveillance systems allows for much quicker targeting of infection hot-spots in pandemic situations, as was done by companies such as Google, in the recent influenza H1N1 crisis (Chew and Eysenbach, 2010; Signorini et al., 2011; St Louis and Zorlu, 2012).

Although search-term surveillance systems such as GFT are currently best suited to track disease activity in developed countries – the system requires large populations of web-search users in order to be most effective (Carneiro and Mylonakis, 2009) and a robust existing surveillance system to provide data for calibration (Wilson et al., 2009), – retrospective analysis of Google Trend's search frequency for the term 'Ebola', in the developing countries of Guinea, Liberia and Sierra Leone, showed a moderate-to-high correlation with epidemic curves for the outbreak in those countries (Milinovich et al., 2015) suggesting that web-based surveillance systems have the potential to be used as early-warning systems in developing, as well as in developed, countries.

However, systems which mine secondary (e.g. news reports) rather than primary web-based data sources (e.g. search queries) are possibly better suited for disease surveillance in developing countries. Examples of such systems include BioCaster (http://biocaster.nii.ac.jp, last accessed April 2015; Collier et al., 2008), EpiSPIDER (Tolentino et al., 2007; Keller et al., 2009), HealthMap (http://www.healthmap.org, last accessed April 2015; Brownstein et al., 2008; Freifeld et al., 2008; Brownstein et al., 2009; Keller et al., 2009; Brownstein et al., 2010), ProMED-mail (http://www.promedmail.org, last accessed April 2015; Cowen et al., 2006; Tolentino et al., 2007; Zeldenrust et al., 2008) and Canada's Global Public Health Intelligence Network (GPHIN) (Mykhalovskiy and Weir, 2006). The value of such systems for flagging potential health threats is highlighted by the fact that GPHIN identified the 2002 severe acute respiratory syndrome (SARS) outbreak in Guangdong Province, China, more than two months before the World Health Organisation's (WHO) official announcement (Mykhalovskiy and Weir, 2006). Similarly, HealthMap identified news stories reporting a strange fever in Guinea nine days before official notification of the 2014 West Africa Ebola outbreak (Milinovich et al., 2015).

Although the inadequate initial response by the international community to the 2014 Ebola outbreak has been highlighted by some as a failure of Big Data analytical approaches for purposes of early warning (Leetaru, 2014; Milinovich et al., 2015), the fact remains that the primary value of such systems currently lies in their ability to flag events that may warrant further investigation rather than acting as the primary surveillance system (Wilson and Brownstein, 2009; Hartley et al., 2013). As such, although web-based surveillance systems are still a long way from replacing traditional surveillance methods, they provide a useful complement to conventional approaches (Milinovich et al., 2014), to the extent that they have become an important component of the