Big Data Usage Patterns in the Health Care Domain: A Use Case Driven Approach Applied to the Assessment of Vaccination Benefits and Risks

Contribution of the IMIA Primary Healthcare Working Group

H. Liyanage[1], S. de Lusignan[1], S-T. Liaw[2], C. Kuziemsky[3], F. Mold[1], P. Krause[4], D. Fleming[1], S. Jones[1]

[1] Clinical Informatics & Health Outcomes research group, Department of Health Care Policy and Management, University of Surrey, Guildford, Surrey, UK
[2] School of Public Health & Community Medicine, UNSW Medicine Australia, NSW, Australia
[3] Telfer School of Management, University of Ottawa, Ottawa, Ontario, Canada
[4] Department of Computing, University of Surrey, Guildford, Surrey, UK

Summary

Background: Generally benefits and risks of vaccines can be determined from studies carried out as part of regulatory compliance, followed by surveillance of routine data; however there are some rarer and more long term events that require new methods. Big data generated by increasingly affordable personalised computing, and from pervasive computing devices is rapidly growing and low cost, high volume, cloud computing makes the processing of these data inexpensive.

Objective: To describe how big data and related analytical methods might be applied to assess the benefits and risks of vaccines.

Method: We reviewed the literature on the use of big data to improve health, applied to generic vaccine use cases, that illustrate benefits and risks of vaccination. We defined a use case as the interaction between a user and an information system to achieve a goal. We used flu vaccination and pre-school childhood immunisation as exemplars.

Results: We reviewed three big data use cases relevant to assessing vaccine benefits and risks: (i) Big data processing using crowdsourcing, distributed big data processing, and predictive analytics, (ii) Data integration from heterogeneous big data sources, e.g. the increasing range of devices in the "internet of things", and (iii) Real-time monitoring for the direct monitoring of epidemics as well as vaccine effects via social media and other data sources.

Conclusions: Big data raises new ethical dilemmas, though its analysis methods can bring complementary real-time capabilities for monitoring epidemics and assessing vaccine benefit-risk balance.

Keywords

Population surveillance; medical record systems, computerized; information science; immunization; public health

Yearb Med Inform 2014:27-35
http://dx.doi.org/10.15265/IY-2014-0016
Published online August 15, 2014

Big Data for Assessing Vaccination Benefits and Risks: A Use Case Driven Approach

Introduction

The rapidly reducing cost of data storage and the increasing bandwidth of communication networks have enabled large volumes of data to be mobilised easily. Due to the proliferation of social media and cloud computing, the collection, storage, and processing of data has become much easier allowing large datasets to be generated and managed in a cost effective manner [1]. It has been reported that the total data generated in the last two years exceeds that amassed throughout the entire history of the digital age [2]. This extended the boundaries of data processing and introduced a new perspective of data analytics embedded in the concept of "Big Data". A quantitative definition of "Big Data" is difficult as the "big" volume is relative to the time of the definition. Datasets considered to be "Big Data" now, may not be considered so in the very near future, due to technological advancements [3]. This is somewhat in contrast to our conventional approaches of study designs where data requirements are set to answer a specific research question.

Processing Big Data

In terms of the nature of data, big data consists of large bodies of unstructured or raw data which cannot be processed using conventional, largely relational data processing techniques. The complexity of the data is often important in characterising datasets as big data. Big data is characterised by IBM according to four dimensions: volume, velocity, variety, and veracity (Box 1) [4]. Big data processing methods address issues such as data volume and heterogeneity in large datasets [5, 6].

The Big Data Advantage

The literature suggests that cloud-based processing of petabytes of data can be achieved at a fraction of the time and cost when adopting big data processing methods [11]. Whilst the wave of big data technologies will not resolve the plethora of data issues that exist in the health care information ecosystem (the system of processing and filtering data between data recording and utilisation) [7], methods for harnessing big data are considered to offer the potential to take a big step forward in improving the quality and efficiency of health care delivery [8].

Method

1 Overview

We have followed an evidence-based approach for developing use cases. This method has been developed based on a

bottom-up approach for model creation by Regnell et al. [13]. The process started with a rapid literature analysis to select use cases that captured common usages of big data in health care potentially relevant to assess the benefits and risks of vaccines. Publications were identified and key information was extracted, focusing specifically on use cases of big data. These use cases were then evaluated to assess their potentiality to improve research methods of assessing vaccine benefits and risks.

2 Search Strategy

The literature analysis employed is suited to the exploratory reviews of emerging technologies in a given domain. It relies on the fact that there are common usage patterns of techniques in the same domain. Therefore, a sample of actual uses of big data in the domain of interest is sufficient for determining the generic usage patterns (known here as use cases).

We searched PUBMED, Medline, Scopus, Web of Science, and the Cochrane Database for publications related to big data in health care. We focused on health care databases, using simple search strings. An adapted PRISMA flow chart is shown in Fig. 1. We included papers which focused primarily on conceptual or practical uses of big data in health care research. This includes instances where big data has been used to highlight the volume of data. Electronic searches were performed retrospectively to select papers published between January 2005 and December 2013. Search was limited to the English language. The number of articles from all searches totalled 485, with 219 being duplications Due to big data being a relatively new concept, no limits were set/placed regarding the type of papers. As such we included any review and report on this topic.

Abstracts were reviewed for applications of big data methods in health care. Key information from each publication was extracted to form a summary (which we refer to as a use case instance). After the screening process, we grouped the use case instances that have common utilisation patterns of big data. Groups were then analysed separately to build the generalised use cases we describe in the next section.

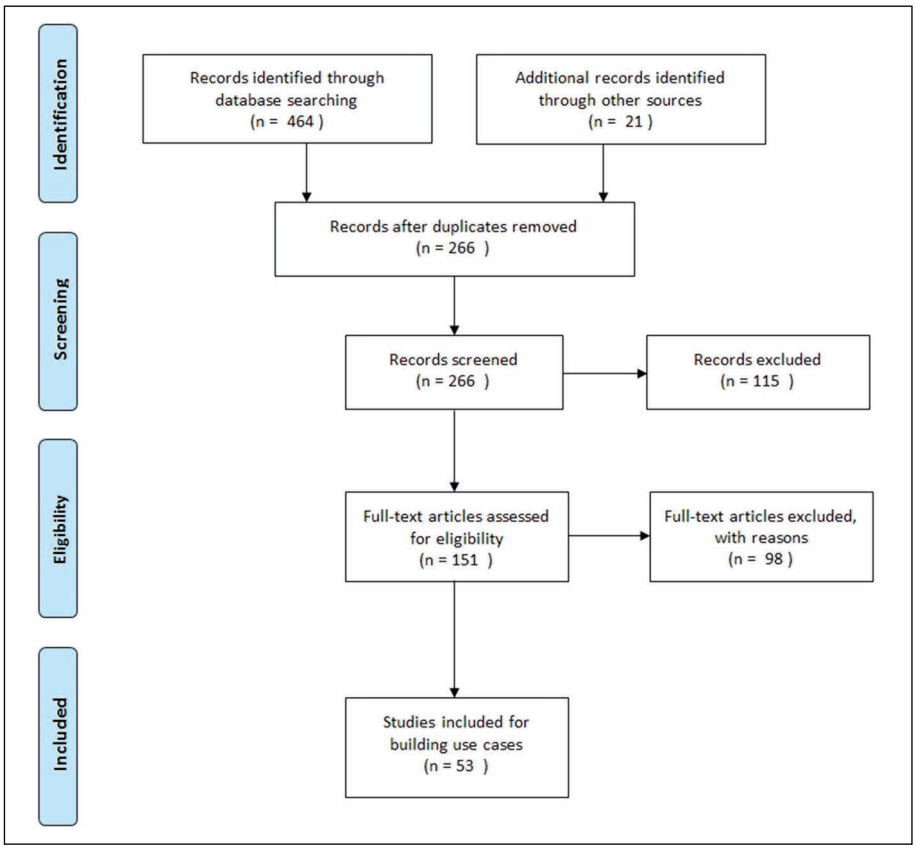

Fig. 1 Adapted PRISMA flow diagram

Table 1 Evidence table used for building the generalised use cases

Generalised use cases	Big data in health care use cases (References)	Vaccine specific use cases
1. Big data processing	Distributed data processing: 28,29,30,31,32 Predictive analysis: 33,34,35,36,37 Crowdsourcing: 38,39	41,42,43,44
2. Data integration of heterogeneous big data sources	45, 46,47,48,49,50,51	52,53,54,55,56,57,58,59,60
3. Real-time monitoring	61,62,63,64,65,66	67,68,69,70,71,72,73,74,75,76,77,78,79

We identified the nature of the dataset (if provided), the field of the study within health care, the particular methods used for manipulating data, and relevance to vaccine research. We used "word clouds" to indicate the key areas where research has been conducted and the methods used [9].

3 Generalised Use Case Models

Use cases are commonly applied in the discipline of software engineering to capture scenarios of how various systems are used in practice [10]. Our group has the experience of developing use cases to support complex research questions [11, 12]. Use cases provide a mechanism of describing the usage

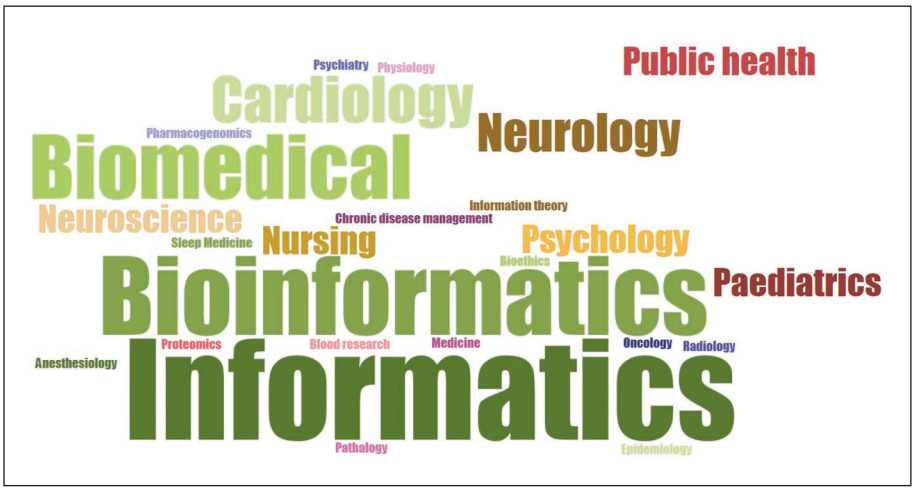

Fig. 2 Word cloud of fields in health care represented by literature search results

of a system and its associated processes. In this approach, specific uses of big data were evaluated and generalised to identify big data usage patterns (i.e. use cases) commonly encountered in the health care domain.

4 Vaccine Benefits and Risks Use Cases

We are involved in a number of programmes that aim to assess the benefits and risks of vaccination within the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) [14,15], which has a long term interest in the monitoring of influenza [16, 17, 18]. We studied the overall benefit-risk for two use-case vaccine programmes: influenza and pre-school childhood immunisation, with a focus on specific adverse events (AE) following immunisation (AEFI) .

This type of research is challenging. For example, it may not be clear whether a flu-like illness is vaccine preventable. Influenza is hard to diagnose precisely and to differentiate from other viral upper respiratory and flu-like illnesses without virology, which is not carried out in routine practice. Flu vaccines are available from a number of sources (e.g. vaccination against flu in the workplace or school), and it can be hard to precisely attribute risk to vaccines; influenza vaccine composition varies every year, and in addition, different vaccine preparations are available each year. Controversy remains about possible long term AEFI, in particular Guillain Barré Syndrome [19, 20]. To date only telephone surveys have been conducted to gain more immediate feedback about local and short term AEFI. Of particular interest is the role of adjuvants [21].

Pre-school childhood vaccinations are usually combination preparations. Making the separation of the responses to vaccination and AEFI is challenging. Passive reporting of AEFI has limitations [22]; and it appears that both patients and medical records under-report febrile seizures that are possible AEFIs [23].

The scope of this paper is to explore how new types of use cases might be developed utilising big data; we will not be addressing these approaches to vaccine research. The purposes of these use cases are threefold:

(1) To demonstrate that by using big data techniques, it is possible to process massive amounts of data that could not be readily processed on a single computer or recruited into a traditional study;

(2) To integrate data that can emanate from a much wider range of observations than would generally be used in medical research; novel sources of data might provide new insights into vaccine benefits and risks. Our focus is on the short and long term, local and systemic AE following influenza vaccination and febrile seizures after pre-school immunisations.

(3) To provide real time monitoring, not daily or weekly, but "as it happens". It may be much faster to report about effectiveness and adverse events. This might provide greater insights into the reactogenicity of different preparations and adjuvants.

Results

The overall result of our analysis was that most studies looked at how big data and associated methods can be leveraged to enhance standard data manipulation methods in different areas of studies in health care. A summary of the big data methods discussed in the publications are highlighted in a word cloud (Fig. 3). In the word cloud, "exploration of data" (including data mining and pattern analysis) was the most emphasised. This is an ubiquitous part of any work with big data, and it is essentially the process of hypothesis generation or early hypothesis testing. We have therefore not created a use case around this activity.

Fig. 3 Word cloud of big data methods represented by literature search results

We have generalised the usage of big data and associated methods into three main use cases (Table 2).

Table 2 Search strings for literature review

Database	Search String
1. PUBMED	"Big data"
2. MEDLINE	"Big data"
3. Scopus	"Big data" AND "health*"
4. Web of science	"Big data" AND "health*"
5. Cochrane	"Big data"

1 Use Case 1: Big Data Processing

Data generation in health care systems has now reached exabyte levels (1 exabyte = 1 billion gigabytes) [24, 25]. Big data technologists have a growing collection of big data processing techniques and we have found evidence that some of these techniques have been used within health care applications. We describe the more widely used big data processing methods.

- ***Distributed big data processing:*** Conventional data processing techniques do not scale to meet processing requirements of big data. MapReduce is a distributed data processing method often used to process big data. This method adopts a two-step approach where the problem is first split into many homogenous sub-problems ('map' step) and then outputs of sub-problems are combined to generate the overall output ('reduce' step) [26]. A cloud-based implementation of MapReduce has been used to analyse electrophysiological data in epilepsy clinical research [27]. The increased sensitivity of cardiac imaging and radiology equipment results in multiple terabytes of data being produced each year and frameworks such as MapReduce have been used for proof-of-concept studies in this area [28]. This processing method is also commonly used in complex biological data processing operations such as genome sequencing which required large computation capabilities [29]. Evidence exists that big data processing outperforms conventional techniques to support the increased size of datasets [30]. Current statistical methods may indeed have limitations when handling the scale of datasets associated with big data.
- ***Predictive analytics:*** This method of analysis uses various statistical and data mining techniques to analyse historical and present data in order to predict future outcomes. Predictive analytics is already demonstrating its usefulness in applications that enable a smarter prediction of health care outcomes by combining clinical, insurance and public datasets. It supports "intelligent case management" which involves the development of programs that can have a higher impact on patient behaviour [31, 32]. We have found evidence for the use of these methods in secondary care for purposes ranging from reducing readmissions to predicting clinical outcomes in patients admitted to surgical intensive care units [33, 34, 35].
- ***Crowdsourcing:*** This involves recruiting large numbers of people who collaboratively collect, filter, and analyse large amounts of data for a common purpose. Using the internet as the medium of participation, thousands of people participate in completing a small part of a problem (often offered in multiplicity for the purpose of validation). Gamified approaches (i.e. regular tasks built as computer games) have been developed for aiding the diagnosis or the labelling of biomedical images [36]. In HIV research, crowdsourcing has been used to identify important genes using 500 billion subsets related to HIV biology [37].

Implications for vaccine research: Vaccines available on the market are effective against only a limited number of circulating strains. Genetic variations of viruses (i.e. antigenic drift) result in regular reformulations of vaccines [38]. This process is costly and time consuming and has a direct impact specifically during pandemic outbreaks. This was evident in the pandemic outbreak of the H1N1 virus (Influenza A (H1N1) virus is the subtype of influenza A virus that is the most common cause of human influenza) where vaccines were made available only after the first wave of the pandemic [39].

Crowdsourcing and cloud computing services which provide a dedicated infrastructure for collecting and processing data are now readily available. At present, any individual with internet access can launch thousands of clusters which could process terabytes of data within a few hours. This would have taken many months and large amounts of money in the prior to utility computing era, in which computational capabilities can be leased on a pay-per-use basis. Cloud services nowadays are configured with the MapReduce framework to facilitate big data processing.

Big data processing methods can accelerate research in vaccines. There have been several proof-of-concept studies for using big data processing methods to improve efficiency in the vaccine development process. A method for conducting phylogenetic analysis has been proposed. This method uses a cloud-based Hadoop framework (an implementation of the MapReduce framework) to analyse evolutionary relationships between viruses [40]. Codon analysis is used in the process of vaccine development and an accelerated method for conducting this analysis using MapReduce aggregation has been proposed [41]. There have been initiatives to use big data for computational modelling of infectious diseases.

Vaccine efficacy trials also generate large volumes of data that could potentially be classified as big data. A vaccine efficacy trial for vaccination against seasonal influenza was conducted during the 2008-2009 season and involved over 43,000 participants. The analysis of the data collected during such trials could benefit from the use of big data methods [42].

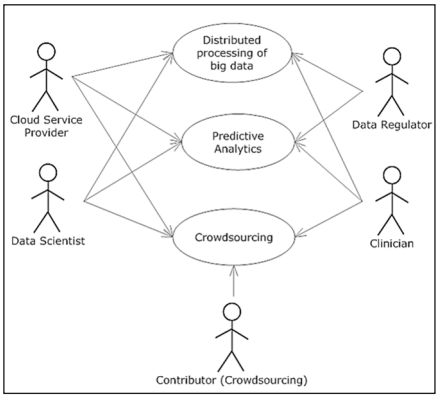

Fig. 4 Big data processing use case diagram

2 Use Case 2: Data Integration of Heterogeneous Big Data Sources

Health care data sources used for decision making are no longer restricted to clinical databases [43]. Accommodating this heterogeneity often leads to better service improvements and policy development. It is important that health care specific techniques such as record linkage, and cohort creation, are adopted appropriately when using big data sources. The majority of the data (80%) in these sources is unstructured and includes medical data such as radiological images, clinic notes, operative reports, and pathologic slides [44, 45]. The national cyber infrastructure of the United States has progressively improved to support the integration of various big data sources, particularly for biomedical research [46]. There also has been an effort to integrate biomedical data sources in Japan [47]. There is evidence of integrating big data sources in studies related to mental health [48]. Examples of data sources used include pervasive device usage data, patient workflows and sensor readings. The Cancer Genome Atlas is a publicly accessible website which acts as a portal to multiple big data sources and includes genomic data, tissue slide images and clinical outcomes data [49]. Evidence demonstrates a stream of work focusing on frameworks developing service-models based on the aggregation of data from pervasive computing devices (wearable body sensors, body function monitors), lifestyle data (diet, sleep), mobility data (smartphones, tablets) and clinical data. Such aggregation helps build an understanding of the overall lifestyle and not only medical conditions, thereby leading to better preventive care.

Implications for vaccine research: This use case emphasises the importance of exploring novel data sources that might provide a better understanding about the effectiveness of vaccines. Most accessible and growing data sources are available on the internet which has resulted in novel concepts such as "Infodemiology" or "Infoveillance". They refer to public health data analysis by using the internet as the main medium which primarily involves aggregating information sources such as social media, websites and web blogs [50].

Social media has proven to be an effective source of unstructured but relevant information. Twitter is frequently used for sentiment studies as the data reflects a large collection of personalised contributions from individuals. A vaccines sentiment study using Twitter data was conducted and network clusters with strong sentiment bias were identified. These associations have allowed the study of individuals and groups in the rich context of their lives, and the study of disease spreading at an individual person level [51]. Although apart from the medical literature, it has been shown during disasters such as the Japan Earthquake or Boston Marathon Bombing that people are moving to social media like Twitter as their main communication tool. Social media tools may become better sources of contextual medical data than patient record systems. Insurance claims data has been utilised as a useful data source for measuring vaccine coverage. It has been effectively used by systems such as PRISM (Post-licensure rapid immunization safety monitoring program) which is associated with the FDA's Mini-sentinel project. This program currently holds data of about 107 million individuals collected through three health insurance organisations [52]. This system conducts an active surveillance through a distributed claims-based database.

Telephone triage data can also be a useful data source, and a proof-of-concept study for surveillance of Influenza-type diseases has been conducted [53].

Specific use case development – Effectiveness of community flu vaccination in older people and Guillain Barré Syndrome as a possible late AE following influenza A vaccination: The focus of how we would harness these technologies for this use case would be to gain further insights into current dilemmas about effectiveness of flu vaccination in the community. Whilst, if the strain is correct, effectiveness in residential homes is proven, this is not known for community transmission [54-57]. A particular research focus would be whether children, and contact with children who might be vectors of disease, will reduce community cases in areas where childhood vaccination is piloted. A recent Cochrane review flags some of the limitations, and possible funding and publication bias of research into childhood vaccines to date [58]. Long term histories and insights into the life events of those with Guillain Barré Syndrome may provide additional insights into this condition, and into what might contribute to individual susceptibility.

3 Use Case 3: Real-time Monitoring

Present day medical devices generate data at a rapid rate which is challenging especially when real-time responses are expected. Similarly, when decisions need to be taken from aggregated data sources such as social media, data needs to be processed in real-time in order to benefit from the relevance of the insights generated from the data. Evidence in

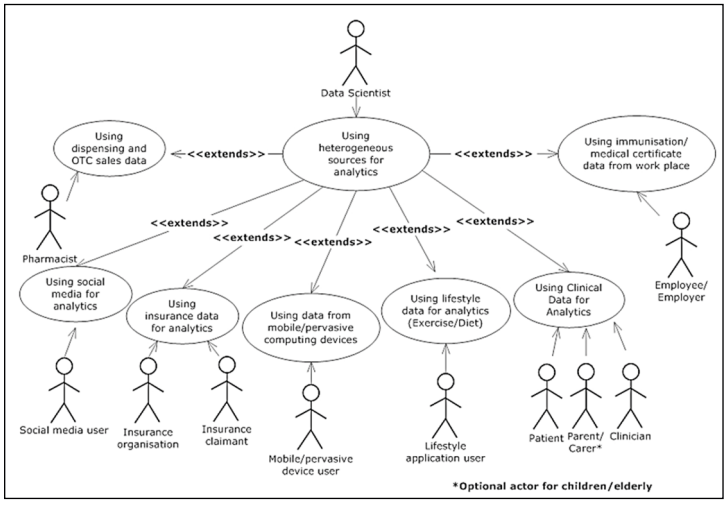

Fig. 5 Data integration of heterogeneous big data sources use case diagram

our review demonstrates real-time monitoring as a significant use case in health care. In pharmacovigilance, once drugs are approved and available in the market, it is required to monitor AEs to ensure that drugs are performing as expected. AEs are generally captured by Adverse Event Reporting Systems and efforts are being made to move towards real-time surveillance of these events. Spontaneous reporting systems (SRS) are currently in use for drug safety monitoring [59].

We have also observed an increasing trend in incorporating social media as a data source for supporting clinical data in health care applications and policy development. The micro blogging service Twitter has so far been the most adopted social media service [60, 61]. Social media data has been proven to be more insightful in assessing patient satisfaction about the quality of care than conventional methods for collecting feedback from forms and surveys [62]. The HealthMap application uses social media data and other online web sources to perform infectious disease surveillance on a global scale [63]. Twitter and similar micro blogging services are increasingly used for public health analysis [64].

Implications for vaccine research: Spontaneous reporting systems have traditionally been used for monitoring vaccine safety. In addition, the Simplified Periodic Safety Update Reports (S-PSUR) and results from observational cohort studies are used as data sources for assessing vaccine safety. There is variability in the availability of exposure data in different countries and there are initiatives to develop multinational automated platforms for achieving this goal [65].

Surveillance of AEFI is important to detect potential vaccine AEs. These AEs may not have been identified during the prelicensure phase and can potentially be serious. The VENICE (Vaccine European New Integrated Collaboration Effort) project was carried out with the intention of harmonising the monitoring of AEs across Europe [66]. With the advent of big data technology there is a trend moving from passive to active surveillance methods and reaching novel data sources including data from screening during hospitalisation, general practice networks, dispensing, over-the-counter sales and social media [67].

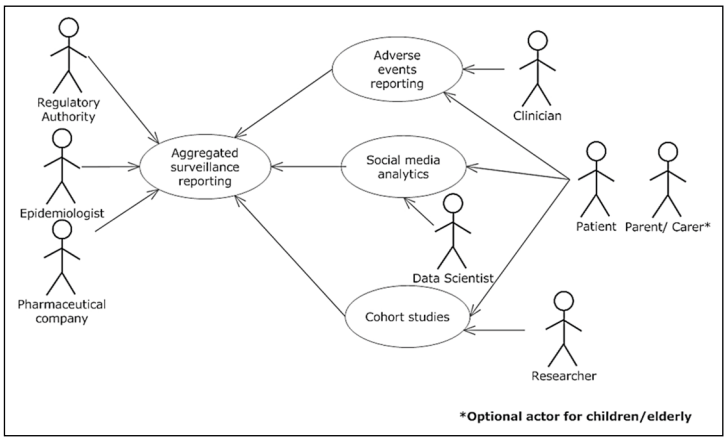

Fig. 6 Real-time monitoring use case diagram

"Google flu trends" uses a novel approach to surveillance in which it uses internet search query data to predict trends in the spreading of infectious diseases [68]. This application has the advantage of being able to access data which cannot usually be integrated by public health programs due to various complexities [69]. This system identifies flu activity about two weeks earlier than conventional systems and is available in more than sixteen countries. Twitter data has been used to predict flu trends in a similar manner [70]. During the Influenza A H1N1 pandemic, Twitter has been used to track levels of disease activity and public concern in the United States [71]. Further evidence in the usage of Twitter for analysing public health indicates the potential for vaccine surveillance [72].

Signal detection is an important aspect of vaccine surveillance. A signal comprises information from one or more sources which suggests a new potentially causal association between an intervention and an event (either adverse or beneficial) [73]. Traditionally signals are detected using the review of individual and aggregated clinical reports. However, this process is rapidly moving towards real-time signal detection through the development of adverse-event monitoring systems [74].

Systems conducting active surveillance will require leveraging non-clinical data sources to effectively detect adverse events. A new wave of biosecurity intelligence systems has effectively used web-based data sources for this purpose [75]. The M-Eco system has successfully trialled the usage of social-media data for signal detection [76].

Specific use case development – Febrile convulsions as a possible AEFI and local reactogenicity: We would explore the development of a specific use case around two areas. Firstly, febrile convulsions as a possible AE following pre-school immunisation, and secondly around local reactivity to immunisations. There are limitations in current reporting systems for AEFI surveillance, they vary in method and by manufacturer [77]. We believe it would be possible to explore rates of febrile convulsions in children, before and after vaccination, and between different surveillance systems. We feel that local reactogenicity would also be amenable to this type of research and it maybe possible to pick up signals about activity and disturbed nights after immunisation. Additionally, video and picture evidence might also be readily collected.

Discussion

1 Principal Findings

There is a growing evidence-base for using big data in health care and this has enabled generalising essential use cases for big data in health care. By using big data approaches, our use cases potentially allow us to more efficiently achieve that goal through processing, integration and monitoring. Current data collection, storage, and analysis approaches are too rigid to support timely vaccination monitoring. This evidence-based approach based on use cases focussed on assessing vaccine benefits and risks. It is likely that all three areas described in these use cases, processing, integration, and real-time monitoring might have a role in future monitoring vaccine benefits and risks.

2 Applying Use Cases to Vaccine Benefits and Risks Monitoring Using a Seasonal Vaccine and a Preventive Vaccine as Exemplars

The benefits and risks of vaccines vary based on the at risk group they are administered to, and according to an individual disease. We use influenza, a seasonal and epidemic infection, as one example and the combined diphtheria tetanus and pertussis triple (DTP3) vaccine given as standard immunisation to children as another. On a worldwide basis, we still know relatively little about the burden of disease from influenza [78, 79]. By extrapolating data from high-income countries, the World Health Organisation (WHO) estimates that annual influenza epidemics result in around 3 to 5 million clinically significant cases and about 300K to 500K deaths worldwide each year. The WHO Global Influenza Surveillance and Response System (GISRS) carries out the virological analysis of around 1 million possible flu samples, though mainly in the developed world. There were about 350 million doses of flu vaccine administered in 2006 rising to around 900 million doses by 2009 [80]. This is a territory for big data methodologies. Global immunisation data provided by the WHO suggest that 111 million children have received the DTP3 vaccine in 2012 [81]. At this scale, this is another ideal candidate for exploring big data technologies.

Use case 1: Big data processing
Crowdsourcing might be a very good way to monitor infectious diseases if self testing or other validated tools could be incorporated into the case detection process. Crowdsourcing might also provide valuable insights into transmission. Distributed data processing and predictive analysis might enhance our ability to predict the spread of epidemics and the emergence of new viral illnesses, and improve the search for factors associated with increased benefits and risks of vaccines.

Use case 2: Data integration of heterogeneous big data sources
Potential benefits for vaccine research may come from the linking of benefits and risks to genotype, biomarkers, and to a whole range of other "infomarkers." These movement data may provide insight into the mechanism of infectious diseases spreading, about which we currently lack a deep understanding, especially from outside the developed world with its largely temperate climate.

Use Case 3: Real-time monitoring
Real-time monitoring considerations might provide insights into any short term side-effects from seasonal immunisation, and enable spatial and temporal patterns of AEs to be described. However, real-time monitoring may have a greater role if it allows to highlight emerging epidemics and disease spread around normal social activities – such as the start of the school term (which is associated with a rise in viral illness in the RCGP RSC), Christmas (where there is usually a lull), and from travels, or purchases of relevant over-the-counter medications [15, 82-84].

3 Implications of Findings

The publicly available health data has the capability to improve the current methods of assessing benefits and risks of vaccines. This would be contingent on the demonstration that data applications used elsewhere in health care could be applied to vaccines. There is certainly an interest for more real-time applications to monitor the effectiveness of vaccines due the capabilities of big data processing methods and the lower cost of utility computer services on the cloud. In order to get the maximum benefit of big data infrastructures, vaccine-specific data processing methods need to be optimised. This may start with the real-time collection of data from people who are also part of current WHO (and/or regional) surveillance systems. This will be challenging to future-proof as there is likely to be an exponential growth of data generated from connected devices and evolving systems such as the Internet of Things (IoT) in the near future. The IoT is being explored as a mechanism for disaster monitoring [85]. Such applications may well be applicable to monitoring epidemics

4 Ethical Aspects of Using Big Data

As the adoption of big data increases, there are increased concerns about the ethical use of health data. Privacy and security needs to be considered as a primary consideration of any big data solution in health care, and the necessary legislation needs to be adopted to ensure that big data is not misused. Effort should be taken to keep individual identities from being identified during big data processing workflows. However, there is the possibility that this may result in data duplication during the integration of big data sources and affect the usefulness of big data.

Legislation in the USA such as HIPAA (Health Insurance Portability and Accountability Act) do not cover data storage outside of health care systems. Much of these big data will not be personal data and not subject to European legislation, and control may remain under the responsibility and ethics of system designers and owners [86]. However, this is a critical concern for those looking to use big data generated and available from online sources. Community driven health data repositories may not be as private as consumers assume [87]. The ownership of data generated through big data analytics is also constantly subjected to debate. Most online services are free and user agreements generally state that the owner of the service can use the data collected from the application. It is necessary to regulate data available online (especially health-related data). Intermediate processors that enforce governance restrictions may be an effective method to handle privacy and ethical concerns of big data [88].

5 Limitations of the Method

The main limitation of this method is that the search did not exhaustively review data sources that may be categorised as big data but not explicitly labelled as such. However, we feel we have identified a sufficient sample of big data scenarios to capture the essential use cases.

Conclusions

Big data is a promising advance that has the capability of improving care, and specifically help the assessment of the benefits and risks of vaccines beyond what can be achieved by just using data from a limited number of health care sources. There are still many issues related to integration and interoperability of data that we have not addressed.

Ethical and governance issues about big data methods, especially in health care, are still to be resolved. This new paradigm of handling big data requires a mind shift in order to effectively leverage the technologies. There is a requirement for novel methods to be developed to make the most of big data. Notwithstanding these limitations, monitoring of vaccines benefits and risks is one area where big data methods, by removing previous limitations on data volumes, sources, immediacy, and processing power, can benefit.

Acknowledgements
This paper is an academic exercise by the Primary Health Care Informatics Working Group of IMIA. This paper and its contributions is the work of volunteers.

References

1. Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation computer systems 2009;25(6):599-616.
2. IBM - Bringing big data to the enterprise - What is big data? - Australia. 2013. IBM - Bringing big data to the enterprise - What is big data? - Australia. [ONLINE] Available at: http://www-01.ibm.com/software/au/data/bigdata/. [Accessed 29 December 2013].
3. Ward JS, Barker A. Undefined By Data: A Survey of Big Data Definitions. 2013;1309.5821.
4. IBM. The Big Data and Analytics Hub. URL: http://www.ibmbigdatahub.com/infographic/four-vs-big-data
5. Dong X, Bahroos N, Sadhu E, Jackson T, Chukhman M, Johnson R, Boyd A, Hynes D. Leverage Hadoop Framework for Large Scale Clinical Informatics Applications. AMIA Summits Transl Sci Proc 2013;2013:53.
6. Demchenko Y, Grosso P, de Laat C. Membrey, P. Addressing Big Data Issues in Scientific Data Infrastructure. Proceedings of International Conference on Collaboration Technologies and Systems 2013;48-55.
7. Neff G. Why Big Data Won't Cure Us. Big Data 2013;1(3):117-23.
8. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13):1351-2.
9. Wordle Word Clouds. URL: http://www.wordle.net
10. Leffingwell D, Widrig D. Managing software requirements: a use case approach. Addison-Wesley Professional; 2003.
11. de Lusignan S, Cashman J, Poh N, Michalakidis G, Mason A, Desombre T, et al. Conducting requirements analyses for research using routinely collected health data: a model driven approach. Stud Health Technol Inform 2012;180:1105-7.
12. Leppenwell E, de Lusignan S, Vicente MT, Michalakidis G, Krause P, Thompson S, et al. Developing a survey instrument to assess the readiness of primary care data, genetic and disease registries to conduct linked research: TRANSFoRm International Research Readiness (TIRRE) survey instrument. Inform Prim Care 2012;20(3):207-16.
13. Regnell B, Andersson M, Bergstrand J. A hierarchical use case model with graphical representation. In: Engineering of Computer-Based Systems, 1996. Proceedings., IEEE Symposium and Workshop. IEEE 1997;270-7.
14. Fleming DM, Miles J. The representativeness of sentinel practice networks. J Public Health (Oxf) 2010;32(1):90-6.
15. Fleming DM, Elliot AJ. Lessons from 40 years' surveillance of influenza in England and Wales. Epidemiol Infect 2008 Jul;136(7):866-75.
16. Pebody R, Andrews N, McMenamin J, Durnall H, Ellis J, Thompson CI, et al. Vaccine effectiveness of 2011/12 trivalent seasonal influenza vaccine in preventing laboratory-confirmed influenza in primary care in the United Kingdom: evidence of waning intra-seasonal protection. Euro Surveill 2013 Jan 31;18(5).
17. Fleming DM, Andrews NJ, Ellis JS, Bermingham A, Sebastianpillai P, Elliot AJ, et al. Estimating influenza vaccine effectiveness using routinely collected laboratory data. J Epidemiol Community Health 2010 Dec;64(12):1062-7.
18. Fleming DM, Verlander NQ, Elliot AJ, Zhao H, Gelb D, Jehring D, et al. An assessment of the effect of statin use on the incidence of acute respiratory infections in England during winters 1998-1999 to 2005-2006. Epidemiol Infect 2010 Sep;138(9):1281-8.
19. Polakowski LL, Sandhu SK, Martin DB, Ball R, Macurdy TE, Franks RL, et al. Chart-confirmed guillain-barre syndrome after 2009 H1N1 influenza vaccination among the Medicare population, 2009-2010. Am J Epidemiol 2013;178(6):962-73.
20. Dodd CN, Romio SA, Black S, Vellozzi C, Andrews N, Sturkenboom M, et al; Global H1N1 GBS Consortium. International collaboration to assess the risk of Guillain Barré Syndrome following Influenza A (H1N1) 2009 monovalent vaccines. Vaccine 2013 Sep 13;31(40):4448-58.
21. Huang WT, Chang CH, Peng MC. Telephone monitoring of adverse events during an MF59®-adjuvanted H5N1 influenza vaccination campaign in Taiwan. Hum Vaccin Immunother. 2014 Jan;10(1):100-3.
22. Parrella A, Gold M, Braunack-Mayer A, Baghurst P, Marshall H. Consumer reporting of adverse events following immunization (AEFI): Identifying predictors of reporting an AEFI. Hum Vaccin Immunother 2014 Jan 9;10(3).
23. Ackerson BK, Sy LS, Yao JF, Craig Cheetham T, Espinosa-Rydman AM, Jones TL, et al. Agreement between medical record and parent report for evaluation of childhood febrile seizures. Vaccine 2013 Jun 12;31(27):2904-9.
24. Newman HB, Ellisman MH, Orcutt JA. Data-intensive e-science frontier research. Communications of the ACM 2003:46(11):68-77.
25. Müller H, Hanbury A, Al Shorbaji N. Health information search to deal with the exploding amount of health information produced. Methods Inf Med 2012 Dec 4;51(6):516-8.
26. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet 2010 Sep;11(9):647-57.
27. Sahoo SS, Jayapandian C, Garg G, Kaffashi F, Chung S, Bozorgi A, et al. Heart beats in the cloud: distributed analysis of electrophysiological 'big data' using cloud computing for epilepsy clinical research. J Am Med Inform Assoc 2014 Mar-Apr;21(2):263-71.
28. Narula J. Are We Up to Speed?: From Big Data to Rich Insights in CV Imaging for a Hyperconnected World. JACC Cardiovasc Imaging 2013 Nov;6(11):1222-4.
29. Mudunuri US, Khouja M, Repetski S, Venkataraman G, Che A, Luke BT, et al. Knowledge and Theme Discovery across Very Large Biological Data Sets Using Distributed Queries: A Prototype Combining Unstructured and Structured Data. 2013 Dec 2;8(12):e80503.
30. Simpson SE, Madigan D, Zorych I, Schuemie MJ, Ryan PB, Suchard MA. Multiple self-controlled case series for large-scale longitudinal observational databases. Biometrics 2013 Dec;69(4):893-902.
31. Fox B. Using big data for big impact. How predictive modeling can affect patient outcomes. Health Manag Technol 2012 Jan;33(1):32.
32. Fox B. Using big data for big impact. Leveraging data and analytics provides the foundation for rethinking how to impact patient behavior. Health Manag Technol 2011 Nov;32(11):16.
33. de Lissovoy G. Big data meets the electronic medical record: a commentary on "identifying patients at increased risk for unplanned readmission". Med Care 2013 Sep;51(9):759-60.
34. Choi M, Lee J, Ahn MJ, Kim Y. Nursing critical patient severity classification system predicts outcomes in patients admitted to surgical intensive care units: use of data from clinical data repository. Stud Health Technol Inform 2013;192:1063.
35. Celi LA, Mark RG, Stone DJ, Montgomery RA. "Big data" in the intensive care unit. Closing the data loop. Am J Respir Crit Care Med 2013 Jun 1;187(11):1157-60.
36. Mavandadi S, Dimitrov S, Feng S, Yu F, Yu R, Sikora U, et al. Crowd-sourced BioGames: managing the big data problem for next-generation lab-on-a-chip platforms. Lab Chip 2012 Oct 21;12(20):4102-6.
37. Bushman FD, Barton S, Bailey A, Greig C, Malani N, Bandyopadhyay S, et al. Bringing it all together: big data and HIV research. AIDS 2013 Mar 13;27(5):835-8.
38. Krammer F, Palese P. Universal influenza virus vaccines: need for clinical trials. Nat Immunol 2013 Dec 18;15(1):3-5.
39. Racaniello, V. Virology blog. http://www.virology.ws/2010/12/09/pandemic-influenza-vaccine-was-too-late-in-2009/ (9 December 2010).
40. Hung CL, Lin CY. Open reading frame phylogenetic analysis on the cloud. Int J Genomics 2013;2013:614923.
41. Radenski A, Ehwerhemuepha L. Speeding-up codon analysis on the cloud with local MapReduce aggregation. Information Sciences2013
42. Eggleston M, Grefenstette J, Burke D. Using big data for computational modeling of infectious

diseases. In: 141st APHA Annual Meeting, 2013; 276179. URL: https://apha.confex.com/apha/141am/webprogram/Paper276179.html.
43. Liyanage H, Liaw ST, Kuziemsky C, Terry AL, Jones S, Soler JK, et al. The Evidence-base for Using Ontologies and Semantic Integration Methodologies to Support Integrated Chronic Disease Management in Primary and Ambulatory Care: Realist Review. Contribution of the IMIA Primary Health Care Informatics WG.Yearb Med Inform 2013;8(1):147-54.
44. Seth Grimes, "Unstructured Data and the 80 Percent Rule: Investigating the 80%". Clarabridge, Bridgepoints; 2008 Q3.
45. Green DE, Rapp EJ. Can big data lead us to big savings? Radiographics 2013 May;33(3):859-60.
46. Leduc R, Vaughn M, Fonner JM, Sullivan M, Williams JG, Blood PD, et al. Leveraging the national cyberinfrastructure for biomedical research. J Am Med Inform Assoc 2014 Mar-Apr;21(2):195-9.
47. Morita M, Igarashi Y, Ito M, Chen YA, Nagao C, Sakaguchi Y, et al. Sagace: a web-based search engine for biomedical databases in Japan. BMC Res Notes 2012 Oct 31;5:604.
48. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral Intervention Technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013 Jul-Aug;35(4):332-8.
49. Robbins DE, Grüneberg A, Deus HF, Tanik MM, Almeida JS. A self-updating road map of The Cancer Genome Atlas. Bioinformatics 2013 May 15;29(10):1333-40.
50. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11.
51. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011 Oct;7(10):e1002199.
52. Baker MA, Nguyen M, Cole DV, Lee GM, Lieu TA. Post-licensure rapid immunization safety monitoring program (PRISM) data characterization. Vaccine 2013 Dec 30;31 Suppl 10:K98-K112.
53. Espino JU, Hogan WR, Wagner MM. Telephone triage: a timely data source for surveillance of influenza-like diseases. AMIA Annu Symp Proc 2003:215-9.
54. Gross PA, Hermogenes AW, Sacks HS, Lau J, Levandowski RA. The efficacy of influenza vaccine in elderly persons. A meta-analysis and review of the literature. Ann Intern Med 1995;123(7):518-27.
55. Jefferson T, Rivetti D, Rivetti A, Rudin M, Di Pietrantonj C, Demicheli V. Efficacy and effectiveness of influenza vaccines in elderly people: a systematic review. Lancet 2005;366(9492):1165-74.
56. Rivetti D, Jefferson T, Thomas R, Rudin M, Rivetti A, Di Pietrantonj C, et al. Vaccines for preventing influenza in the elderly. Cochrane Database Syst Rev 2006 Jul 19;(3):CD004876.
57. Jefferson T, Di Pietrantonj C, Al-Ansary LA, Ferroni E, Thorning S, Thomas RE. Vaccines for preventing influenza in the elderly. Cochrane Database Syst Rev 2010 Feb 17;(2):CD004876.
58. Jefferson T, Rivetti A, Di Pietrantonj C, Demicheli V, Ferroni E. Vaccines for preventing influenza in healthy children. Cochrane Database Syst Rev 2012 Aug 15;8:CD004879.
59. Shah NH. Translational bioinformatics embraces big data. Yearb Med Inform 2012;7(1):130-4.
60. Bartlett C, Wurtz R. Twitter and Public Health. J Public Health Manag Pract 2013 Dec 18. [Epub ahead of print]
61. Denecke K, Krieck M, Otrusina L, Smrz P, Dolog P, Nejdl W, et al. How to exploit twitter for public health monitoring? Methods Inf Med 2013;52(4):326-39.
62. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Qual Saf 2013 Mar;22(3):251-5.
63. Hay SI, George DB, Moyes CL, Brownstein JS. Big data opportunities for global infectious disease surveillance. PLoS Med 2013;10(4):e1001413.
64. Paul MJ, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. ICWSM 2011 July;
65. Eurosurveillance editorial team. ECDC in collaboration with the VAESCO consortium to develop a complementary tool for vaccine safety monitoring in Europe. Euro Surveill. 2009 Oct 1;14(39). pii: 19345.
66. Zanoni G, Berra P, Lucchi I, Ferro A, O'Flanagan D, Levy-Bruhl Det al. Vaccine adverse event monitoring systems across the European Union countries: time for unifying efforts. Vaccine 2009 May 26;27(25-26):3376-84.
67. Crawford NW, Clothier H, Hodgson K, Selvaraj G, Easton ML, Buttery JP. Active surveillance for adverse events following immunization. Expert Rev Vaccines 2013 Dec 18.
68. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009;49(10):1557-64.
69. Jalali A, Olabode OA, Bell CM. Leveraging Cloud Computing to Address Public Health Disparities: An Analysis of the SPHPS. Online J Public Health Inform 2012;4(3).
70. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Predicting flu trends using twitter data. In: Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference 2011. p. 702-7.
71. Signorini A, Segre AM, Polgreen PM. The use of twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic. PLoS One 2011;6: e19467.
72. Love B, Himelboim I, Holton A, Stewart K. Twitter as a source of vaccination information: Content drivers and what they are saying. Am J Infect Control 2013:41(6):568-70.
73. CIOMS. Practical Aspects of Signal Detection in Pharmacovigilance. CIOMS; 2010,
74. Iskander JK, Miller ER, Chen RT. The role of the Vaccine Adverse Event Reporting system (VAERS) in monitoring vaccine safety. Pediatr Ann 2004 Sep;33(9):599-606.
75. Lyon A, Nunn M, Grossel G, Burgman M. Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. Transbound Emerg Dis 2012 Jun;59(3):223-32.
76. Denecke K, Krieck M, Otrusina L, Smrz P, Dolog P, Nejdl W, et al. How to exploit twitter for public health monitoring? Methods Inf Med 2013;52(4):326-39.
77. Guo B, Page A, Wang H, Taylor R, McIntyre P. Systematic review of reporting rates of adverse events following immunization: an international comparison of post-marketing surveillance programs with reference to China. Vaccine 2013 Jan 11;31(4):603-17.
78. Simonsen L, Spreeuwenberg P, Lustig R, Taylor RJ, Fleming DM, Kroneman M, et al; GLaMOR Collaborating Teams. Global mortality estimates for the 2009 Influenza Pandemic from the GLaMOR project: a modeling study. PLoS Med 2013 Nov;10(11):e1001558.
79. Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng PY, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. Lancet Infect Dis 2012 Sep;12(9):687-95. Erratum in: Lancet Infect Dis 2012 Sep;12(9):655.
80. World Health Organisation (WHO). Report of the second WHO consultation on the global action plan for influenza vaccines (GAP), Geneva, Switzerland, 12–14 July 2011. URL: http://whqlibdoc.who.int/publications/2012/9789241564410_eng.pdf
81. WHO | Immunization surveillance, assessment and monitoring. 2014. WHO | Immunization surveillance, assessment and monitoring. [ONLINE] Available at: http://www.who.int/immunization/monitoring_surveillance/en/. [Accessed 05 December 2013].
82. Dailey L, Watkins RE, Plant AJ. Timeliness of data sources used for influenza surveillance. J Am Med Inform Assoc 2007 Sep-Oct;14(5):626-31.
83. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proc Natl Acad Sci USA 2002 Apr 16;99(8):5237-40.
84. Ohkusa Y, Shigematsu M, Taniguchi K, Okabe N. Experimental surveillance using data on sales of over-the-counter medications--Japan, November 2003-April 2004.MMWR Morb Mortal Wkly Rep 2005 Aug 26;54 Suppl:47-52.
85. Yang L, Yang SH, Plotnick L. How the Internet of things technology enhances emergency response. Technological Forecasting And Social Change 2013;80(9):1854-67.
86. de Lusignan S, Chan T, Theadom A, Dhoul N. The roles of policy and professionalism in the protection of processed clinical data: a literature review. Int J Med Inform 2007 Apr;76(4):261-8.
87. Steinbrook R. Personally controlled online health data-the next big thing in medical care? N Engl J Med 2008 Apr 17;358(16):1653-6.
88. Liyanage H, Liaw ST, de Lusignan S. Accelerating the development of an information ecosystem in health care, by stimulating the growth of safe intermediate processing of health information (IPHI). Inform Prim Care 2012;20(2):81-6.

Correspondence to:
Simon de Lusignan
Clinical Informatics & Health Outcomes research group
Department of Health Care Policy and Management
University of Surrey
GUILDFORD
Surrey GU2 7XH, UK
E-mail: s.lusignan@surrey.ac.uk