PharmacoEconomics (2016) 34:101–106
DOI 10.1007/s40273-015-0306-7

CURRENT OPINION

Big Data and Health Economics: Strengths, Weaknesses, Opportunities and Threats

Brendan Collins[1]

Published online: 21 June 2015

Abstract 'Big data' is the collective name for the increasing capacity of information systems to collect and store large volumes of data, which are often unstructured and time stamped, and to analyse these data by using regression and other statistical techniques. This is a review of the potential applications of big data and health economics, using a SWOT (strengths, weaknesses, opportunities, threats) approach. In health economics, large pseudonymized databases, such as the planned care.data programme in the UK, have the potential to increase understanding of how drugs work in the real world, taking into account adherence, co-morbidities, interactions and side effects. This 'real-world evidence' has applications in individualized medicine. More routine and larger-scale cost and outcomes data collection will make health economic analyses more disease specific and population specific but may require new skill sets. There is potential for biomonitoring and lifestyle data to inform health economic analyses and public health policy.

Key Points for Decision Makers

'Big data' technologies have the potential to advance health economics as a discipline.

Having larger patient datasets will allow a lot more real-world evidence to be generated and interactions between treatments to be understood better.

Having more biomonitoring and lifestyle data will enable health interventions to be tailored more to individuals.

1 Introduction

The objective of this paper is to provide the reader with some ideas around developments that have been collectively labelled as 'big data' and how they might have an impact on the field of health economics in the years ahead. This piece uses a SWOT analysis approach to understanding the strengths, weaknesses, opportunities and threats of big-data techniques in the fields of health economics (including pharmacoeconomics), epidemiology and public health. As with any SWOT analysis, there is an element of subjectivity behind the ideas in this paper. The paper first defines and introduces big data (talking specifically about open data), talks about the opportunities for health economics and biomonitoring, and then discusses large data repositories, the opportunities for public health, and some of the risks and challenges around big data.

There are several definitions of big data; most definitions talk about unstructured, large and often time-stamped datasets, which would be difficult to process in standard

✉ Brendan Collins
brenc@liv.ac.uk

[1] University of Liverpool Management School, Chatham Street, Liverpool L69 7ZH, UK

relational databases [1]. Big data involves technologies that allow unprecedented opportunities to store, match up, analyse and visualize datasets in ways that would not have been possible in the past, and which can reveal aspects of human behaviour and processes that were previously difficult to measure. There is a question over whether big data is a genuinely new phenomenon. There are big datasets that have been used for years, such as the National Health and Nutrition Examination Survey (NHANES) in the USA; the volumes of digital data surpassed the volumes of non-digital data globally in 2002 and now constitute 94 % of all data. Business intelligence technology has moved quickly, but the real big-data revolution is in the volumes of data that are collected and stored (with the amount of information stored globally being estimated to double every 40 months), and the increase in the processing power of computers and servers, with people now talking about peta-, exa-, or zettabytes, where in the past they would refer to kilobytes. We now collect data without necessarily knowing the purpose for collecting them. In the past, developers would spend months trying to get information systems to speak to each other, whereas now it is usually a lot quicker and automated. In the past, data were used to measure against a performance or a plan, or to test hypotheses, whereas data mining and automated neural network models are now used to generate new hypotheses and to look for links that may not immediately be logical or apparent—'unknown unknowns'. In the last 3 years, it has been revealed that the National Security Agency (NSA) in the USA and the Government Communications Headquarters (GCHQ) in the UK have been monitoring people's phones and internet use on a massive scale—action that is possible only through big-data storage and sifting technology. There is a debate around whether big data will enhance our human potential or will mean being over-monitored and manipulated without our knowledge.

Big data often involves 'open' datasets, which are shared in the public domain. In future, amateur or volunteer 'citizen scientists' may compete with academic groups; for example, PXE International has found treatments for a rare disease, pseudoxanthoma elasticum (PXE), by sharing data and blood samples with interested parties [2]. The UK Government has had a programme of putting data into the public domain and has a central repository, data.gov.uk, for these data; for example, one sector of the UK Government with a lot of potential applications is Ordnance Survey (OS), which has allowed a lot of its previously paid-for geographical mapping products to be used openly [3]. In terms of health, the UK Government has faced negative headlines and a backlash over its plans for care.data, a web interface where linked general practitioner (GP) and hospital admissions data would be available to researchers, potentially creating one of the world's biggest linked health datasets. Much of this negative publicity has arisen because of the possibility of data being shared with private companies [4]. In the USA, the Patient-Centered Outcomes Research Institute (PCORI) has funded similar big-data projects [5]. These types of datasets could be crucial in understanding the whole patient journey and meeting healthcare challenges such as efficiency savings or moving services out of hospitals.

Phil Hammond, a British doctor, broadcaster and campaigner, has called for open data in healthcare and aggressive transparency [6], and for the UK health service to do more to protect whistleblowers who expose poor standards of care. But having open, transparent data could put healthcare givers at greater risk of litigation when things go wrong. It is often quoted that around 1 in 10 people is harmed by healthcare, although this estimate varies widely [7]. There are risks that with open data, people will miss the nuances and misinterpret; for example, in the case of US Medicare cost data being released, some physicians were identified as being the biggest earners when in fact they were managing health budgets for large programmes [8].

There are also 'closed datasets', which are owned and sold by private companies, such as Quintiles. These closed datasets also have increased potential in big-data applications through data linkage. Dr. Foster, a company that was initially set up in partnership with the UK Department of Health, analyses closed datasets to provide hospital data analytics, including the controversial Hospital Standardised Mortality Ratios (HSMRs), which were part of identifying the high mortality rates in the Mid Staffordshire NHS [National Health Service] Trust, which led to the Francis Inquiry. These mortality ratios are adjusted for case mix, age and co-morbidities so that a fair comparison can be made, although this is disputed by some who argue that it is over-sensitive to local differences in clinical coding within hospitals [9]. Hospital readmissions are notoriously tricky to predict, despite there being several algorithms already, such as PARR+ (Patients at Risk of Readmission), that estimate a patient's probability of being readmitted. The Heritage Health Prize was a competition where analysts were given access to a large hospital training dataset and were asked to produce an algorithm that would predict hospital readmissions as accurately as possible in a test dataset [10].

Table 1 shows a list of the types of datasets and an example of each one. These datasets are not all uniquely linked to the big-data phenomenon; some have been around for many years. But within each category, big data will mean that capability is increased.

Most big-data applications use the internet. The influential book *Super Crunchers* [11] detailed how companies

Table 1 Types of health-related data that big data will have an impact on

Type of health-related data	Example
Healthcare registries	MINAP
Healthcare claims databases	Medicare
Adverse events databases	FAERS
Lifestyle databases	Experian Mosaic
Biomonitoring data	Propeller Health (asthma devices database)
Large population surveys	NHANES
Internet browsing/searching databases	Google Flu Trends
'Internet of Things'	Thingful

FAERS US Food and Drug Administration's Adverse Events Reporting System, *MINAP* [UK] Myocardial Ischaemia National Audit Project, *NHANES* [US] National Health and Nutrition Examination Survey

such as Amazon record every click a customer makes and how long it takes, and sometimes they test out different prices for the same product; this has been employed controversially by budget airlines, who increase prices when a customer looks at a flight several times. Healthcare data are one of the types of data that individuals are less willing or less likely to share on the internet, so it could be argued that there are fewer applications for acquiring these data routinely. Google Flu Trends was a big-data 'collective intelligence' system, which recorded web searches for flu symptoms and was postulated as an early warning system for flu outbreaks. However, despite some early success, it was found in subsequent analyses to be not very accurate in predicting rates of flu cases [12]. Because Google's algorithms are proprietary and not openly available, they cannot be interrogated by other researchers.

2 Strengths of Big Data and Opportunities for Health Economics

Big-data technologies offer opportunities for health economics and pharmacoepidemiology in terms of exponentially larger datasets, linking between systems and using real-world evidence to see how drugs interact in the real world and to identify safety issues and rare side effects more quickly. Big data offers the possibility for a lot more complexity to be measured and stored in the healthcare system, which previously may not have been possible with slow information systems that could not talk to each other. But big data probably will not replace the gold standards of clinical research—for instance, having a hypothesis and running a randomized, controlled trial. In the UK NHS, increasing use of service-line reporting of healthcare spending and routine measurement of health outcomes in programmes such as the national Patient Reported Outcome Measures (PROMs) programme mean that potential data for cost-effectiveness analyses may be captured routinely in hospital databases as opposed to requiring additional data collection and planning; for example, a study used PROMs data for a cost–utility analysis of hip and knee replacement [13]. Large datasets may also allow increased monitoring of consumer behaviour in the healthcare system to more accurately determine whether health insurance coverage improves health, increases healthcare usage or creates moral hazard, such as that in the famous RAND health insurance experiment and, more recently, the Oregon experiment [14].

3 Biomonitoring and Lifestyle Data

The increased use of passive biomonitoring devices (such as continuous heart rate monitors) will mean that in future, more real-time data on health indicators will be available and healthy ranges for biometrics can be better defined. This is part of a movement towards what has been called 'lifelogging' or the 'quantified self', where more individual data are collected, covering health, lifestyle and daily activities. In May 2014, Samsung unveiled a commercial health sensor called Simband [15]. This device measures metrics such as the heart rate, blood pressure and oxygen levels. However, the level of accuracy needed for a commercial device may not be rigorous enough for a health study. Increasing lifestyle data from activity monitoring, from shopping habits and from the 'Internet of Things' (which is defined as a future scenario where household functions such as heating, fridges and doors will be connected to the internet and controlled remotely) can be used to inform policies to improve public health through increasing activity and access to healthy foods [16]. Data from these novel sources may present issues in the short term—for instance, where institutional review boards (IRBs) are not used to assess whether and how researchers should have access to these types of data sources from human subjects. Research

Table 2 Strengths, weaknesses, opportunities and threats of big-data technologies and methodologies for the pharmaceutical industry and health economics, and for the general public

	For the field of health economics	For the general public
Strengths	Open datasets mean that more robust long-term outcomes data are available for economic models. Also, more accurate disaggregated cost data are available that represent the whole patient journey Open datasets mean that people from different disciplines and people from outside the research industry can analyse data and test hypotheses (‘citizen science’), bringing fresh perspectives and opportunities for more open collaboration Real-world evidence means that the cost effectiveness of drugs in the real world can be better understood (taking into account the effects of multiple drugs, and the fact that adherence is often lower in the real world than in trials), which may provide better evidence for decision makers Availability of more individual-level data means that the effects of drugs at the individual level and changes in outcomes and value to individuals can be better understood	Real-world evidence should mean more tailored, better drugs and greater efficiency, which means that the system is more efficient for insurance payers and taxpayers Interactions arising from uncommon drug combinations should be better understood Individual preferences can be better measured so that the human consequences of treatments can be incorporated into decisions in a more precise way. Thus, if a drug helps somebody to live independently or to continue to engage in pleasurable activities, these changes in capability can be factored in for cases where there may be only a small improvement in typical health outcome measures, such as the EQ-5D Giving healthcare providers and patients access to relevant, well presented, high-quality data should help them to make better decisions Behavioural data from sensors and the ‘Internet of Things’ can be used to make communities healthier by promoting healthy behaviours such as walking and cycling, and fruit and vegetable consumption
Weaknesses	It can be costly to store and manipulate large volumes of data Spurious associations could be identified when multiple analyses are used without being sense-checked Does not meet the level of scientific rigour needed in randomized, controlled trials and thus will not replace them	Big data favours people who are more digitally connected, i.e. those in rich countries People might feel that they are losing their individual sovereignty if they are being over-monitored, such as by biomonitoring systems People do not always know what their data are being used for, thus they may be less likely to participate in trials
Opportunities	Have large datasets that communicate with each other better, which can facilitate complex analyses Have large trial registries so that the effects of similar drugs can be cross-matched Possibility of understanding mechanisms and side effects of new drugs better through big datasets, which should reduce the number of new drugs that fail Potential to further personalize medicines by informing how drugs and other treatments work in individual patients, leading to improved efficiency Capacity to build large model datasets that resemble real populations in size and complexity	Individuals can understand their genetic predisposition to diseases Individuals can use their own biomonitoring data to obtain better and more immediate treatment, or to monitor and improve their own health More accurate predictions of new developments in diseases means quicker responses, e.g. in flu epidemics Availability of more accurate predictive data means better prevention of diseases, e.g. for understanding of cancer risk factors
Threats	Real-world evidence might show that drugs are not as effective in the real world Health economics as an industry might be required less if more data collection, analysis and modelling can be automated The health economics field might suffer a backlash in the public consciousness if people feel that their data are being misused or that they are being over-monitored Health economics as a discipline may need to develop a new skill set for working with big-data applications	Understanding a lot more about individual risks of diseases may cause anxiety, particularly for diseases with no treatments available Risk of adverse selection by insurance companies with better predictive datasets Risk of companies using big datasets to collude at the expense of the customer Risk of data being lost or stolen Risk of biological indicators of health being valued over individual welfare Ethical risks around excessive genetic screening

protocols may take time to catch up with new methodological and ethical challenges arising as a result of access to these novel data sources.

Table 2 shows a summary of the strengths, weaknesses, opportunities and threats of the big-data approach for the field of health economics and for the general public.

4 Real-World Evidence and Personalized Medicine

Health economists have broadly welcomed the move towards having big data available and generating more ‘real-world evidence’, even though some companies might not benefit in cases where drugs are shown to be less effective in the real world than they were in clinical trials. Drug companies already use consultancies to trawl clinical data systems to see how effective their drugs are in the real world (and also arguably to check which clinicians are prescribing their products). Big-data techniques chime with ideas around personalized medicine, whereby clinicians can select treatments that are most effective and less likely to be discontinued because of side effects, and, in the case of drugs, they can select a dose that is tailored more to an individual, potentially making the healthcare system much more efficient.

5 Data Repositories

The ‘All Trials’ initiative, which several big pharmaceutical companies have signed up to, is calling for all clinical trials to be reported in an open central repository, and could represent a big-data resource. The NHS Economic Evaluation Database (EED) and the Cost-Effectiveness Analysis (CEA) Registry are popular databases of economic evaluations, while the Cochrane Collaboration is a well-respected database of systematic reviews. The US Food and Drug Administration’s Adverse Events Reporting System (FAERS) is a system for reporting adverse events and medication errors for drug companies, clinicians and the public, which has big-data applications [17]. If these registries could be connected with planned databases of trials, then there may be applications such as automated systematic reviews and greater sharing between drug companies and research groups. There may be ethical questions because traditionally in trials, people are told what their data are being used for, whereas with big data, the uses of their data may not always be known yet. However, some big datasets, such as NHANES and the Health Survey for England, have been studied for many different purposes over the years [18].

One example of where big-data techniques could help to settle a health controversy is around the prescribing of statins to otherwise healthy people. There has been recent debate about the efficacy of statins, which are prescribed to reduce cholesterol and reduce cardiovascular risk [19]. In particular, there is debate about the correct threshold of cardiovascular risk at which to prescribe statins, as there is also a small risk of side effects such as diabetes. In the UK, the National Institute for Health and Care Excellence (NICE) has recommended that statins should be considered for adults who have a 20 % or greater 10-year risk of developing CVD [20]. A lot of the current cardiovascular risk equations, such as the Framingham Risk Equation and QRisk, have limitations [21]. Big data could enable systems to be linked up so that more sophisticated estimates of the correct threshold could be produced at an individual patient level, along with their level of certainty. Then, if a risk calculation can be embedded into the health system, there could be a regression equation that keeps testing its reliability and updating itself to become more powerful.

6 Modelling

In health economics, modelling is often used to combine data from different sources or to estimate long-term outcomes from short-term surrogate outcomes. Data on quality of life and costs are often reused between many studies and are often used for populations that are different from those being studied, and such data may be outdated. If data collection, processing and analysis in economic studies can become more sophisticated and efficient, then a greater volume of accurate healthcare outcomes and costing data can be generated. It is traditional for cost–utility analyses to take an extra-welfarist perspective, where the value of health is based on population averages [22]. For example, in the UK, the utility scores associated with particular health states in the EQ-5D (which is seen as the gold standard by NICE in calculating quality-adjusted life expectancy) are calculated using time trade-off methods and a general population sample. This is partly so that health investment decisions reflect the views of taxpayers, but it is limited by the ability of people to “put themselves in someone else’s shoes” and imagine health scenarios. With big-data applications, it may be that the individual value that people put on health states can be used more often, and outcomes such as the capability of an individual to do the things they enjoy could be incorporated, which would represent moving towards more of a welfarist perspective. In terms of modelling, big data means that in future, modelled populations or simulations could be closer to real populations in their size and complexity. However, the opportunities around using big-data techniques will not compensate for a model that has been poorly designed.

7 Risks of Big Data and Threats to Health Economics

There are risks to using big data, such as loss of patient confidentiality or misuse of data by insurers or other companies. In the US ‘Obamacare’ healthcare reforms, health insurance markets were not allowed to consider pre-existing conditions in pricing policies for individuals [23], because it was recognized that companies will soon be able to predict healthcare costs with accuracy by using big-data applications. In *Privacy in the Age of Big Data*, Payton and Claypoole stated that the biggest threat to individuals is from cyber-criminals, as most databases can be hacked, even secure government databases such as the Pentagon's [24]. There is also a risk that the benefits of biomonitoring and the Internet of Things will leave the most deprived populations behind. There have been rapid advances in genomics; a company called 23andMe previously offered genetic testing which told individuals their risks of certain diseases [25]. Genetic testing such as this could pose ethical and moral questions, especially for diseases where there are few effective treatments, such as Huntington's chorea.

8 Summary

For health economists, big-data technologies may mean that more data can be collected and combined from disparate information sources and with automated analysis. In the future, open-ended research data may be used more to generate, as well as to test, hypotheses. Real-world evidence may indicate that drugs do not perform as well in the real world as they do in randomized, controlled trials, but, equally, personalized medicine may provide evidence of benefit for individual patients who may be considered on the borderline of whether a given treatment would be cost effective. The increasing use of big data will offer benefits to individuals, such as more tailored healthcare solutions, better monitoring of health conditions and fewer mistakes. But there will also be risks of patients feeling too much like they are a series of numbers and less like individuals, and risks of data being stolen.

In conclusion, the big-data revolution is a real phenomenon and will mean real changes in the field of health economics, and having an appreciation of the risks and benefits of it is useful for practitioners in the field. There may be an increasing demand for analytical skills in working with large datasets in health economics, as well as in other research fields. There should be opportunities in making the health system more efficient and in tailoring drugs and information more to individuals.

Compliance with ethical standards The author received no specific funding for this study and has no conflicts of interest to declare.

References

1. Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc. 2012;15(5):662–79.
2. Solovitch S. The citizen scientists. Wired Magazine. 2001;9. Available at: http://archive.wired.com/wired/archive/9.09/disease.html. Accessed 16 June 2015.
3. Turnbull G. How to find all the maps published on GOV.UK. Government Data Service blog. 2014. Available at: http://gds.blog.gov.uk/2014/02/14/how-to-find-all-the-maps-published-on-gov-uk/. Accessed 1 May 2014.
4. Sheather J, Brannan S. Patient confidentiality in a time of care.-data. BMJ. 2013;347:f7042.
5. Menius JA Jr, Rousculp MD. Growth in health care data causing an evolution in the pharmaceutical industry. NC Med J. 2014;75(3):188–90.
6. Berger A. TV: trust me. I'm a surgeon. BMJ. 2000;320(7239):948.
7. The Health Foundation. Levels of harm: evidence scan. 2011. Available at: http://www.health.org.uk/public/cms/75/76/313/2593/Levels%20of%20harm.pdf?realName=PYiXMz.pdf. Accessed 31 May 2014.
8. Rosenbaum R. What big data can't tell us about healthcare. The New Yorker. 2014. Available at: http://www.newyorker.com/online/blogs/currency/2014/04/the-medicare-data-dump-and-the-cost-of-care.html. Accessed 31 May 2014.
9. Hawkes N. Patient coding and the ratings game. BMJ. 2010;340:950–2.
10. El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, Gluck J. De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. J Med Internet Res. 2012;14(1):e33. doi:10.2196/jmir.2001.
11. Ayres I. Super crunchers: how anything can be predicted. London: John Murray; 2007.
12. Lazer DM, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science. 2014;343(6176):1203.
13. Jenkins PJ, Clement ND, Hamilton DF, Gaston P, Patton JT, Howie CR. Predicting the cost-effectiveness of total hip and knee replacement: a health economic analysis. Bone Jt J. 2013;95(1):115–21.
14. Baicker K, Taubman SL, Allen HL, Bernstein M, Gruber JH, Newhouse JP, Finkelstein AN. The Oregon experiment—effects of medicaid on clinical outcomes. N Engl J Med. 2013;368(18):1713–22.
15. Fitzsimmons M. Simband is Samsung's wearable sensor platform for all. Tech Radar. 2014. Available at: http://www.techradar.com/news/portable-devices/simband-is-samsung-s-open-platform-wearable-platform-1250844. Accessed 31 May 2014.
16. Chui M, Löffler M, Roberts R. The internet of things. McKinsey Q. 2010;2:1–9.
17. Sakaeda T, Tamon A, Kadoyama K, Okuno Y. Data mining of the public version of the FDA adverse event reporting system. Int J Med Sci. 2013;10(7):796.
18. Staa TPV, Goldacre B, Gulliford M, Cassell J, Pirmohamed M, Taweel A, Delaney B, Smeeth L. Pragmatic randomised trials using routine electronic health records: putting them to the test. BMJ. 2012;344:e55.
19. Abramson JD, Rosenberg HD, Jewell N, Wright JM. Should people at low risk of cardiovascular disease take a statin? BMJ. 2013;347:f6123.
20. Cooper A, O'Flynn N. Guidelines: risk assessment and lipid modification for primary and secondary prevention of cardiovascular disease: summary of NICE guidance. BMJ. 2008;336(7655):1246.
21. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRisk cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. Heart. 2008;94(1):34–9.
22. Brouwer WB, Culyer AJ, Van Exel N, Rutten FF. Welfarism vs. extra-welfarism. J Health Econ. 2008;27(2):325–38.
23. McDonough JE. The road ahead for the Affordable Care Act. N Engl J Med. 2012;367(3):199–201.
24. Payton TM, Claypoole T. Privacy in the age of big data: recognizing threats, defending your rights, and protecting your family. Lanham: Rowman & Littlefield; 2014.
25. Annas GJ, Elias S. 23andMe and the FDA. N Engl J Med. 2014;370(11):985–8.