Big Health Data: Institutional and Technological Challenges

Justin Keen [1], Radu Calinescu [2], Richard Paige [2] and John Rooksby [3]

[1] Leeds Institute of Health Sciences, University of Leeds
[2] Department of Computer Science, University of York
[3] School of Computer Science, University of St Andrews

Introduction

Much of the enthusiasm about Big Data focuses on the potential for exploiting the large, high quality datasets that are now available about the weather, sub-atomic particles, voting behaviour and many other topics. Current challenges focus on exploiting the datasets – on the management of data, on new methods for analysis and visualisation. But there is another class of Big Data, typified by the routine data that are generated within organisations, including public services. Datasets are typically inaccurate and incomplete, and yet contain the best available information about the numbers of people who are in the criminal justice system, are doing well at school or are at risk of developing a particular disease.

This paper focuses on the second class of Big Data, and in particular on secondary uses of health care datasets in the NHS in England. For the purposes of discussion, the primary use of health care information is for the diagnosis, treatment and care that we receive from doctors, nurses and other clinicians. Every other use is a secondary use, and includes planning services, managing budgets, accounting for resources used and research. There is a long history of these secondary uses in all health care systems, and citizens appear to be relaxed about the use of aggregated – and non-identifiable – data by staff working within a system. But expectations have been raised about new uses of health care data by third parties (ie neither patients nor health care staff). Manyika and

colleagues (2011), for example, assert that the benefits of Big Health Data could run into the hundreds of billions of dollars, realised through a combination of re-engineering health services and commercial exploitation. We ask two questions - what kind of political thinking informs Big Health Data policies, and what are their prospects? We suggest that it is an example of neo-liberal thinking, though with the dice loaded in favour of firms rather than individuals participating in free markets. We also conclude that, at least at the moment, the expectations for Big Health Data are misplaced. In common with information technology policies in recent times, Big Data policies are based on top-down and abstract ideas, which do not take account of the organisational or technological realities on the ground (Margetts, 6 and Hood, 2010).

In the next section we briefly outline some of the early uses of large health care datasets, noting recent trends in linking systems and datasets, and extending the scope of data being collected. Then we set out recent Government policies on Open Data in England, which we believe will substantially shape Big Data policies and practices, and identify the Government's strategic objectives in this area. We comment on the politics of Big Health Data, noting that current policies are consistent with Harvey's arguments about 'neoliberalism in practice'. In order to evaluate the prospects of success we identify three challenges – data management, technology infrastructure and information governance – and discuss the extent to which each of these is likely to support or undermine achievement of the Government's objectives.

Background

Health systems around the world have long collected large volumes of data, and had to develop practical solutions to the problems of managing them. Mainframe systems were first used for administrative purposes in the 1960's, notably for counting (eg how many patients had been admitted to a hospital) and for managing hospital appointments. They were also used to manage early population-wide systems for research, and for auditing illnesses and deaths, for example in the Framingham Heart Study in the USA and in the Nottingham Heart Attack Register in England. IT systems were gradually introduced into more and more areas of health care in the 1970's and 1980's. By the end of the

1990's almost all GPs in England routinely used computers in their clinical work, and most hospital departments had systems that captured activity information (who had been born, who had been operated on, who had had which pathology tests and so on).

A snapshot taken at the turn of the millennium would have revealed large numbers of systems and datasets, used for a wide range of purposes. Relatively few of the 'live' systems were, however, linked to one another: a clinician on a ward who wanted to know the details of a patient's surgery did not have access to the hospital's operating theatre system, or to that patient's GP records. Large national and regional datasets for secondary use, such as registries for cancer and heart disease, were managed by different organisations, often far removed organisationally from frontline clinicians.

These developments were sometimes encouraged by the Department of Health – the growth of GP computing is a good example – but in general they occurred without much central direction. The NHS had IT policies in the 1990's, but there was little money to influence implementation. The situation changed dramatically in 2002, when the Labour administration launched the NHS National Programme for IT, designed to computerise every aspect of NHS services and management, and hold data in (what was in effect) a single logical server for the whole Service. Large, long-term contracts for systems were agreed with consortia of suppliers. The Programme was doomed from the start. Its objectives were never clear, there was little experimental evidence that the chosen technologies would be cost-effective, the technologies were not proven, and the politics were naïve, with civil servants attempting to impose solutions on powerful groups of professionals. (Indeed, there is some evidence that Prime Minister Blair and senior civil servants based the decision to go ahead on unexamined beliefs about the inherent value of information technologies in health care (Brennan 2005, Nicholson 2008)). The story of the Programme has passed into folklore: key systems were not delivered, large firms pulled out of the Programme or were fired. It is the largest civilian IT implementation failure, anywhere, to date (Keen 2010, National Audit Office 2011).

On the ground, when NHS organisations realised that the Programme would not deliver, they began to implement systems themselves, just as they had done up to 2002. In the last few years hitherto separate systems have been progressively linked together, both within and between organisations, with the technology strategy varying from place to place. Viewed from above, the developments have been federal, or bottom-up in nature. In some places datasets have been linked using patient identifiers, with the aim of enabling doctors and other clinicians to view all of the available information about a patient in a single integrated record. In others, portals have been developed, allowing clinicians to view several systems, remotely, from a single terminal – without data integration. Today, some services are still not extensively computerised, notably for the management of medicines in hospitals and for the support of community nurses: these 'gaps' are gradually being filled in, so that the coverage of systems is also increasing.

It is one of the curiosities of the NHS that IT policies have historically been separate from data policies. In the last decade there has also been a marked increase in the number of datasets that NHS organisations have collected, and/or use to provide information to the Department of Health. The nature and purpose of the datasets are various, but by way of example include length of hospital stay (evaluated against a main target of 18 weeks from GP referral to clinical action), GP data to evaluate their performance in relation to their national employment contracts (in the Quality and Outcomes Framework), new disease registers for diabetes and stroke, and a system for reporting adverse events to the National Patient Safety Agency. The net effect of these new datasets has been centralising, in two senses of the term: the Department of Health has defined what data are to be collected, and the principal users appear to be the Department and national agencies such as the NHS Information Centre and the Care Quality Commission.

We are not aware of any detailed surveys of the uses of these datasets. Our subjective judgement is that there is extensive use of datasets collected for performance management purposes – in spite of 20 years of market-oriented reforms the NHS is still a large bureaucracy – and highly variable use of datasets for the review of services for people with particular conditions, or quality and safety. We are on firmer ground when

we observe that major failures, such as unacceptably poor quality and safety of services in hospitals, were not identified using routine data – even though the data were available. We will argue later that there are problems with many of these datasets, not least that they are chronically incomplete, and this must limit their usefulness. But it seems reasonable to say that, overall, large NHS datasets are not being fully exploited.

The developments within the NHS are taking place within two much broader trends. One involves the – relatively slow and painful – encounter between the computing and medical devices industries. Hitherto free-standing medical devices are gradually being linked into networks, raising the possibility that data from devices can be pushed to any location, and linked to data from other devices or from information systems. While hospitals have traditionally been home to sensors, of many kinds, they are increasingly being promoted for use in peoples' homes, so that individuals' health status (eg their blood sugar or respiration) can be monitored remotely. If one accepts UK Government estimates that up to three million people might benefit from these home-based devices, then there is clearly potential for generating large volumes of new data.

The second trend is in the area of genomics. It is now possible to analyse an individual's DNA and other genetic material at relatively low cost, and in a matter of hours. The prospects for identifying susceptibility to diseases, and for identifying new strategies for treating diseases with a genetic component, have been hyped for at least the last decade. As Topol (2012) points out, however, there have been few breakthroughs that can be used to improve diagnosis and treatment. (The most promising area to date has been in the area of pharmacogenomics, where our understanding of who might benefit from particular drugs, and who is unlikely to, is improving). This said, the Big Data trend is in one direction: more people are likely to have more genetic information in their records in years to come.

From Big Technology to Open Data

The Coalition Government has continued with Labour's commitment to greater use of markets and competition in health care in England, including the use of non-NHS

organisations to provide services. It has also retained, but re-designed, the performance management infrastructure, focusing on patient experiences and health outcomes as well as on cost and activity measures. The status of the NHS National Programme is still unclear, but it is currently impossible politically for the Government to argue for large scale IT investments.

It is, though, still possible to pursue policies which are heavily dependent on those IT investments. Open Data policies, being pursued across government, seem likely to shape Big Data activities substantially in the next few years: as we have already seen, large scale data collection activities across the NHS have always been centrally driven. This will in turn influence IT investment decisions: individual NHS organisations will have to consider central data requirements when they specify local systems. In November 2011 the Chancellor of the Exchequer, George Osborne, presented his Autumn Statement to Parliament. It included the following passage:

"Making more public sector information available will help catalyse new markets and innovative products and services as well as improving standards and transparency in public services. The Government will open up access to core public datasets on transport, weather and health, including giving individuals access to their online GP records by the end of this Parliament. The Government will provide up to £10 million over five years to establish an Open Data Institute to help industry exploit the opportunities created through release of this data." (Paragraph 1.125)

In a speech in December 2011 the Prime Minister gave more details:

"Now there's something else that we're doing … and that is opening up the vast amounts of data generated in our health service. From this month huge amounts of new data are going to be released online. This is the real world evidence that scientists have been crying out for and we're determined to deliver it…
We're going to consult on actually changing the NHS constitution so that the default setting is for patients' data to be used for research unless of course they want to opt out.

Now let me be clear, this does not threaten privacy, it doesn't mean anyone can look at your health records but it does mean using anonymous data to make new medical breakthroughs and that is something that we should want to see happen right here in our country. Now the end result will be that every willing patient is a research patient; that every time you use the NHS you're playing a part in the fight against disease at home and around the world."

A new NHS IT strategy, published in May 2012, drew attention to changes in legislation. Although centrally driven IT programmes are out of favour, centralisation of data collection is not:

"The Health and Social Care Act 2012 includes provisions marking a step-change in the health and care sector's approach to transparency, growth and open data. It requires the Health and Social Care Information Centre to publish (in safe, 'de-identifiable' format) virtually all of the data it is required to collect across the health and care sector. The Information Centre has already started routinely releasing the data that underpins their statistical publications. As part of this a further 83 datasets were released for the first time in 2011-12, completing the roll-out of this approach.
… The Department understands that knowing which information is available is one of industry's biggest 'asks' of it. To this end, the Act requires the Information Centre to maintain and publish a register ('catalogue') of the data it has collected. In addition the Department will ask the Information Centre to undertake work to develop an inventory of the wealth of data collected by other parts of the health and social care system so that over time it can provide a single source of information on the data that is collected and where it can be accessed.
… In health there are major benefits from linking data – to industry, to research, to providers and commissioners of care services as well as to patients, service users and the broader public – so that we understand more about the whole patient journey, not just isolated episodes of care." (Department of Health 2012, Annex B)

The strategy sketches out the vision for Big Data and Open Data. This includes the planning of services by NHS organisations, commissioning services and research (paras. 3.26-3.29). The strategy also confirms the proposals for the "release of Big Data", the "capture and release of My Data: provision of access for service users to their own identifiable data", and "the creation of dynamic Information Markets" to drive social and economic growth (Annex B). And, it emphasises the importance of IT in supporting individuals' capacity to care for themselves (by giving them access to their GP records), and in enabling them to choose between (competing) hospital services.

Making Sense of Open Data

The statements and the new strategy help us to identify key Open Data policy objectives. These are:

1. NHS performance - better commissioning and performance management of NHS services;
2. To support commercial research and development (R&D), particularly in the pharmaceutical industry;
3. To promote economic growth, through the creation of a new market for health care information;
4. To enable individuals to pursue their own health-related objectives.

What do they tell us about the nature of Open Data policies? We suggest that they are consistent with aspects of neo-liberal thought. Neo-liberal theory emphasises strong individual property rights, freely functioning markets and free trade, and strong legal institutions to underpin the property rights and markets. There are many advocates of neo-liberal thinking and practices, and just as many critics. For our purposes here, Harvey's (2005, pp. 64-86) critique is particularly useful. He argues that, as neo-liberal ideas have entered mainstream policy making in the last 30 or so years, some of the tensions inherent in them have been resolved in particular ways. Two interest us here. First, if tensions arise between individuals and firms, for example over property rights, those tensions tend to be resolved in favour of firms. Second, neo-liberals tend to be

suspicious of government, which *inter alia* tends to impede the operation of free markets. In practice, though, governments have to exist in some form, and it is therefore necessary to integrate state policy making into market processes in some way. Some of the practical ways of achieving this integration are familiar – public-private partnerships, the easy movement between senior civil service positions and large private firms, and regulatory regimes that are favourable to those firms.

Viewed in this light, Open Data policies are usefully viewed as an example of 'neo-liberalism in practice'. The quotes in the last section highlight the importance of property rights – for personal information – and attempt a resolution which is favourable to private firms. They also demonstrate co-operation between the state and private firms, with the former explicitly supporting the aspirations of the latter. Putting the two points together, we can sketch out a set of relationships between four interests, namely the state, private firms, citizens/patients and clinicians. Open Data reinforces relationships between state and private sector actors, and does so by weakening the positions of both citizens/patients and clinicians. We will comment on this point at the end of the next section, in relation to information governance.

Three Challenges

While the Chancellor's and Prime Minister's comments are general in nature, and the information strategy is aspirational rather than detailed, the direction of travel is clear. In this section we return to Big Data, on the basis that the acquisition, storage and manipulation of NHS datasets underpins Open Data and other Government policies. As already noted, substantial claims have been made about the potential of Big Health Data. But in order to realise the hoped-for gains, or anything like them, there are a number of straightforward pre-conditions for success, and we comment on four here, namely data management, technology infrastructure and governance.

Data Management

There are challenges associated with working with Big Data, and in particular with the nature of big health care datasets. Broadly speaking, these challenges are associated with

collection of data, the quality of data, the management of previously collected data, and the destruction of (unwanted) data. We consider each briefly in turn.

Big datasets, like NHS datasets, pose significant collection challenges. A key issue is that such data is often keyed in, which is an expensive and error-prone process. Better automation for supporting data collection is desirable, and there are advances in this area (e.g., better keypads and user-interfaces that are designed to make data entry faster and less error prone), but the challenges of using such advances *at scale* remain poorly understand.

There are also sins of omission in collecting data. Historically, in the NHS, data has been collected along functional lines, and as a result it can be very difficult to use it to evaluate the performance of (parts of) the organisation along key dimensions, such as the effectiveness of coordination of services. For example, a great deal of information is collected about older peoples' use of health services, including their visits to hospital, their blood tests and so on. But information about the management of clinical risks (such as the risk of an older person falling), which is important in preventing or minimising problems, and for the NHS in managing demand, is not routinely collected. *Awareness* of such omissions is difficult to promote and obtain because of the scale and pace of data collection. Promising developments in the field of automatic network scanning (Holm 2012) for populating big datasets and models thereof may offer a capability that reduces the expense and error rate of data collection, though at the price of incompleteness: such techniques can generally only be told what to look for during collection. Combining network scanning with machine learning may offer means to mitigate this.

A second key challenge is the quality of collected data. We have already mentioned the problems associated with manual entry: in general for Big Data, and particularly NHS datasets, quality is inconsistent, particularly in terms of *accuracy* of the collected data. Part of the difficulty in collecting accurate data is due to the rate of data acquisition, and the use of error-prone entry mechanisms, but there are more complex errors due to it not being clear what the datasets will be used for. A dataset is, in effect, a *model* (indeed,

there is even a standard for modelling datasets, the Common Warehouse Metamodel (CWM)), and a model is always constructed for a purpose (e.g., statistical analysis, exploration, explanation). Accuracy of a model is always easier to judge *and improve* when it is clear what the model will be used for; in principle, the purpose of the dataset should inform the collection (and quality assurance) mechanisms used to obtain it, which should further inform any changes made to the intended usages of the dataset. However, big datasets often serve many purposes (e.g., for commissioning services, for statistical analysis by the NHS Information Centre, informing ministerial decisions) and as such it can be difficult, if not impossible, to understand how to improve or judge their accuracy. Finally, there is a significant *freshness* issue associated with big datasets: they become stale over time and identifying when they are stale, and when something must be done about it (e.g., disposal of the dataset, refreshing it, auditing it) is always difficult.

The third key challenge is related to managing big datasets, once data have been collected. There are a number of difficulties here. For some ultra-large datasets that involve complex inter-relationships between entities, standard data management tools (e.g., relational databases) can be difficult to apply; promising work on NoSQL (ie not only SQL) databases may help to address this. Moreover, the procedures and processes that are currently in place for managing small datasets can be difficult to use - particularly for information governance - for big data. A particular challenge relates to audit of access to big datasets, and the granularity of information required to properly audit the accesses. There are also significant challenges related to the efficiency of access and management of big data: such data is simply difficult to store, and many organisations do not have the facilities to do so. Hence, public, private or hybrid cloud storage services will become increasingly important. Finally, there is a difficulty associated with combining big datasets: this will undoubtedly be important to generate new insights (e.g., by combining data from different health datasets), but there may be *emergent* issues that come from the combination: if big datasets have inherent quality (accuracy) concerns, will these concerns be magnified in unpredictable, emergent ways once different datasets are combined?

Our final challenge is something that we just raise as a concern: how and when should big datasets be disposed of? The emphasis today is on data collection and storage. At some point, big datasets will be of limited or no value - certainly not valuable enough to warrant their continued maintenance. When should we decide to dispose of them, and how can we do so in a consistent, secure manner? Current research and practice on data disposal will be invaluable here (Hopkins 2008), but with big data, cross-organisation governance concerns will also come into play.

Technology Infrastructure

From an engineering perspective, the advent of Big Data poses exciting new challenges for technology infrastructure. For Big Data advocates, though, the excitement is a problem, because it stems from the fact that we do not currently know how to solve some important problems. Existing data storage, modelling, querying and analysis paradigms do not directly scale to Big Data, and the current technology support for safe storage and handling is unsuitable for the envisaged uses of these data. The implications of the federated, bottom-up development of NHS IT systems – the same pattern is found in many other countries – become clear here. We can say that technology is playing catch-up with the Government's Big Data vision. The success of new policies will depend, at least in part, on how successfully challenges are addressed by the many research projects that the UK and EU, and administrations around the world, have recently funded in this area.

In terms of data modelling, progress has been made with the development of NHS data standards in the last two decades, but there is still a lack of standards compliance in some services (notably social care), of effective interoperability between systems and of useful metadata to support important trends in health services. This limits the possibilities for automating the linkage of data sets in meaningful ways. For example, the data.gov.uk site provides access to hundreds of health-related datasets, but these are of limited value to practitioners or researchers for these reasons. There is a further problem, which is that some legacy data sets do not have effective schemas, and cannot easily be manipulated, modelled and queried.

Query and analysis of Big Data requires approaches capable of representing and managing data quality, provenance and uncertainty that are just not yet available. Because of the cost and time required to process big data sets, novel analysis paradigms are required that can carry out a partial query and expose its results for a researcher to decide whether to proceed to a full analysis or to discard a spurious research hypothesis. A technological challenge not encountered before the big data era is bringing big datasets and the software that analyses them together (on the same IT infrastructure). The vision here is that we will increasingly encounter applications that require software to "travel" to where the data are located. This paradigm shift from the traditional approach of transferring data to where it is required is needed because transferring vast amounts of data may take too long - or simply because the required storage capacity is not available at the destination.

Big data analysis will require considerable computation power – even by current standards - for potentially short, infrequent periods of time. While this pattern of demand could be addressed through storing and processing Big Data on cloud computing infrastructure, the envisaged move to a G-Cloud (Cabinet Office 2011) would necessitate solving technology challenges concerned with re-architecting enterprise systems for new and very different technology platforms. A key technological challenge here is to resolve the myriad undocumented interdependencies among the numerous information systems that comprise such enterprise systems. The brittle nature of health enterprise systems (Peacock et al 2012) will be stressed by a move to cloud infrastructure, leading to unpredictable timescales and costs.

Discovering relevant data sets in an ocean of big data sets is another substantive challenge. A simple-term search for "mortality"-related data sets on the data.gov.uk site returned 43 data sets (on 25th July 2012), out of which only a few may be relevant for exploring a given research hypothesis. Without dedicated discovery services – such as smart data set directories annotating data sets with key metadata – identifying these few relevant data sets can be very time consuming and costly.

Finally, and anticipating some of the arguments in the next section, information governance rules and policies that accompany the plans to open up Big Health Data are not supported by current technologies. The implications for data protection extend beyond what existing technology can handle, eg privacy protection against statistical inference attacks (ie attacks that reveal sensitive data through statistical analysis of multiple data sources) is an open research question. Although rigorous, established solutions do exist for some security-related aspects of information systems (including, for instance, authentication, access control, and encryption of sensitive data), implementing them correctly is recognized as a challenge. If cloud infrastructure is used to analyse Big Data, powered by new and complex "virtualisation" system software, it is likely to contain security loopholes that will take some time to uncover and fix.

Information Governance

Discussions of information governance in health care typically focus on the privacy and confidentiality of personal information. Privacy and confidentiality are recognised as a major concern across public services in the UK (O'Hara 2011), and by health care policy makers elsewhere (Kundra 2011). The NHS is subject to a wide range of general legislation, including the Data Protection Act 1998. This allows the NHS to collect information for use in diagnosis and treatment of individuals, but explicitly outlaws any other uses, without the consent. As we have already seen, there have long been secondary uses of personal information, but many useful planning and management activities can be undertaken using aggregated, non-personal, information. The NHS Information Centre, and individual NHS organisations, publish highly aggregated data for general consumption, or more detailed datasets only into tightly regulated environments, such as 'safe havens' governed by research ethics agreements.

As we have also seen, though, there are occasions when large datasets retain personal identifiers, notably in the case of disease-based datasets for heart disease, diabetes and other conditions (http://www.hqip.org.uk/national-clinical-audit-registries/). The Health Act 2006 and Health and Social Care Act 2008, between them, outline governance

arrangements for accessing personal information without obtaining the consent of individuals (ie creating an exception to the DPA 1998). Thus planners and researchers can, if they obtain appropriate formal permissions, combine datasets using personal identifiers (eg name, full postcode) without the consent of the identified individuals. The exceptional legal status of health care datasets reinforces the point that they can be used to pursue wider public policy objectives, as well as for the diagnosis and treatment of individual patients.

On the face of it, the legislation creates a promising governance framework for Big Data. In practice, however, there is uncertainty about the nature and extent of Big Data activities (defining this as the use of secondary data), flowing from the Open Data policies outlined earlier. We highlight four issues here. First, current regulations do not provide a basis for navigating in the new legislative environment, eg by providing guidance on dealing with the health-related risks of Privacy 2.0 (Zittrain 2008), where hackers can combine data from a range of sources to create detailed profiles of individuals, beyond the reach of any regulator. These need to be developed and implemented before publication of large numbers of datasets are published.

Second, we have argued that the Government and commercial interests are asserting property rights over personal health care information. As Manson and O'Neill point out (2007), this sort of arrangement can only work if we, as users of the NHS, trust the institutions that manage our information. If we trust the NHS and third parties, then we may be relaxed about the publication of relatively detailed datasets, and if not then not. The Prime Minister has, at least implicitly, taken the view that we will not trust the relevant institutions, by saying that personal information will be anonymised. The obvious question is: if records really are anonymised, how useful are they to any firm? A detailed discussion of this point is beyond the scope of this paper, but we note the general point, that there is a relationship between trust in institutions and the granularity of published information.

Third, Big Data enthusiasts underestimate, or just ignore, the social processes involved in defining the data to be collected, and in data collection and use in practice. Star (1999; Star and Ruhleder 1996; Star and Bowker 2002) has argued that infrastructure ought not to be viewed just in terms of the technology, but in terms of the ways in which it is produced and sustained across heterogeneous sites and over the long term. Infrastructure, particularly "information infrastructure" (Anderson et al 2008), will be an amalgamation of technologies, data and practices produced under varying conditions across a variety of sites. Therefore, infrastructure relies as much on cooperation, organization and trust (Lee and Dourish 2006; Jirotka et al 2005; Bietz et al 2010) as it does on having technologies and standards in place. Star and Rhudler (1996) point to the social practices and institutions within which infrastructures are embedded: the transparency or taken-for granted-ness of infrastructure (except when it breaks); its spatial and temporal reach across sites and uses; its co-evolution with the uses it is put to; its relation with standards; and its relation with an already installed base of other infrastructures and technologies. Star and Rhudler point out that "nobody is really in charge of infrastructure", which is not to say it isn't designed or managed, but that it is developed and changed through negotiation, and often in piecemeal and evolutionary ways. Star draws attention away from infrastructure as a noun and to the work of "infrastructuring", the ongoing processes of creating and sustaining infrastructure. Experience in eScience, for example in the establishment of metadata standards (Bietz et al 2010) and the establishment of standard terminologies (Hine 2008), has shown that creating infrastructure can be extremely problematic as it involves shifts in relations, responsibilities and status between organisations. The consequences of such infrastructuring will often be unclear and there will be many uncertainties and risks. The outcomes may also suit some more than others.

The fourth and final governance issue concerns the relationship between Open Data and other NHS policies. While it is possible to imagine circumstances where publication of datasets will support other policies, for example in providing information about local services, it may also undermine them. One important current example is competition. The more health care organisations such as NHS Trusts and commissioning bodies are encouraged to compete, the more reticent they are likely to be to release information

about their performance. Competitors will be interested in information about the service delivery model that a Trust is using – maybe it has made Lean Production work in health care – or in the costs of delivering those services. The Coalition Government may, therefore, find itself having to decide whether Open Data trumps competition, or *vice versa*.

Conclusions

We have addressed two questions in this paper – what kind of political thinking informs Big Health Data policies, and what are their prospects? We have suggested that the use of large secondary datasets will be substantially shaped by Open Data policies, and that these policies are usefully viewed as neo-liberal in nature. We have further suggested that success or failure will turn on the question of trust in institutions. If trust is lacking, as the Prime Minister seems to fear, then the NHS is likely to adopt cautious publication policies, and it is a moot point whether published datasets will be of interest to firms, either in the desired information market or to boost to pharmaceutical R&D.

The arguments in the last section highlight the substantial practical obstacles to the achievement of the Coalition Government's objectives. There is a long way to go to provide appropriate data and technology infrastructures, and information governance arrangements that provide citizens/patients with assurances that their personal information will be properly protected. A realistic assessment of the current state of affairs is a necessary pre-condition for making progress towards the Government's objectives.

References

Anderson S, G Hardstone, R Procter, R Williams. "Down in the (Data)base(ment): Supporting Configuration in Organizational Information Systems." In *Resources, Co-Evolution and Artifacts*, 221–253. London: Springer.

Bietz M, E Baumer, C Lee. 2010. "Synergizing in Cyberinfrastructure Development." *Computer Supported Cooperative Work (CSCW)* 19 (3): 245–281. doi:10.1007/s10606-010-9114-y.

Bowker G, Star SL. Sorting Things Out. Cambridge MA, MIT Press, 1999.

Brennan S. 2005. An interview with Sir John Pattison, in S. Brennan, *The NHS IT Project*, pp. 190-4. Oxford, Radcliffe.

Cabinet Office. *Cloud Computing Strategy*. http://bit.ly/vFSQ28

Cameron D. *Speech on Life Sciences and Opening Up the NHS*, 6 December 2011. http://bit.ly/s4hXEG

Department of Health. *The Power of Information*. London, Department of Health.

Hannes Holm, Markus Buschle, Robert Lagerström and Mathias Ekstedt, Automatic data collection for enterprise architecture models, accepted and to appear in *Software and Systems Modeling,* Springer-Verlag, 2012.

Harvey D. 2005. *A Brief History of Neoliberalism*. Oxford, Oxford University Press.

Hine, Christine. 2008. *Systematics as Cyberscience: Computers, Change, and Continuity in Science*. The MIT Press.

Hopkins R, K Jenkins. 2008. *Eating the IT Elephant*. IBM Press.

Jirotka, Marina, Rob Procter, Mark Hartswood, Roger Slack, Andrew Simpson, Catelijne Coopmans, Chris Hinds, and Alex Voss. 2005. "Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study." *Computer Supported Cooperative Work (CSCW)* 14 (4) (September 14): 369–398. doi:10.1007/s10606-005-9001-0.

Keen J. 2010. Integration at any Price. In: Margetts H, 6 P, Hood C. *Paradoxes of Modernization*. Oxford, Oxford University Press, 2010.

Lee C, P Dourish, G Mark. 2006. "The Human Infrastructure of Cyberinfrastructure." In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 483–492. CSCW '06. New York, NY, USA: ACM. doi:10.1145/1180875.1180950. http://doi.acm.org/10.1145/1180875.1180950.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A. Big Data: The Next Frontier for Innovation, Competition and Productivity. McKinsey Global Institute, May 2011.

Manson N, O O'Neill. 2007. *Rethinking Informed Consent in Bioethics*. Cambridge, Cambridge University Press.

Margetts H, 6 P, Hood C. *Paradoxes of Modernization*. Oxford, Oxford University Press, 2010.

National Audit Office. 2011. The National Programme for IT in the NHS: an update on the delivery of detailed care records systems. HC888, Session 2010-12. London, TSO.

Nicholson D. 2008. *Oral evidence in Public Accounts Committee*. The National Programme for IT in the NHS: Progress. HC153, Ev. 1-20. London, TSO.

O'Hara, K. Transparent government, not transparent citizens. http://bit.ly/oyo3po

Kundra, V. http://bit.ly/otcopn

Peacock R, Moore J, Keen J. Interim Realities: Information Technologies and Unknown Destinations. *Public Management Review* 2012.

DOI: 10.1080/14719037.2012.657836

Star S. 1999. "The Ethnography of Infrastructure." *American Behavioural Scientist* 43 (3): 377–391.

Topol E. 2012. The Creative Destruction of Medicine: How the Digital Revolution Will Deliver Better Health Care. Basic.

Zittrain J. The Future of the Internet. London, Penguin, 2008.