Research Article

Keen / Calinescu / Paige / Rooksby: The Case of Health Care Data in England

Big Data + Politics = Open Data: The Case of Health Care Data in England

Justin Keen, University of Leeds [j.keen@leeds.ac.uk]
Radu Calinescu, University of York
Richard Paige, University of York
John Rooksby, University of Glasgow

There is a great deal of enthusiasm about the prospects for the Big Data held in health care systems around the world. Health care appears to offer the ideal combination of circumstances for its exploitation, with a need to improve productivity on the one hand and the availability of data that can be used to identify opportunities for improvement on the other. The enthusiasm rests on two assumptions. First, that the datasets held by hospitals and other organizations, and the technological infrastructure needed for their acquisition, storage and manipulation, are up to the task. Second, that organizations outside health care systems will be able to access detailed datasets. We argue that both assumptions can be challenged. The article uses the example of the National Health Service in England to identify data, technology, and information governance challenges. The public acceptability of third party access to detailed health care datasets is, at best, unclear.

KEY WORDS: Big Data, Open Data, health care, information technologies

Acknowledgments: This paper reports on research funded by the UK Engineering and Physical Sciences Research Council, through its Large Scale Complex IT Systems Programme, EP/F001096/1 and EP/H042644/1.

Introduction

There is a great deal of enthusiasm about the prospects for the Big Data held in health care systems around the world. Health care appears to offer the ideal combination of circumstances for its exploitation, with a need to improve productivity on the one hand and the availability of data that can be used to identify opportunities for productivity improvements on the other. Data have historically been held in paper records, or in isolated systems, but many countries have moved from paper to electronic records and begun to link systems in recent years. On the face of it, it should now be possible to produce datasets for monitoring performance, for research, and other purposes.

The enthusiasm rests on two assumptions. The first is that the datasets held by hospitals and other organizations, and the technological infrastructure needed for their acquisition, storage and manipulation, are up to the task. Data need to be relevant and complete, and the infrastructure needs to support the secure storage and manipulation of large volumes of data. The second assumption takes us into the realm of Open Data, the notion that data currently held by public bodies should be published, and hence available to all of us to use. One argument for Open Data is that public services, including health services, generate large volumes of data that they do not exploit effectively. (It could also be argued that we pay for the data, through our taxes and insurance, and should therefore have access to it.) If datasets can be published, in forms that protect individuals' confidentiality, third parties will be able to use it effectively. Advocates argue that any large dataset is potentially valuable, though the greatest opportunities for commercial opportunities probably lie in detailed person-level datasets. Manyika and colleagues (2011), for example, assert that the benefits of Big Health Data

could run into the hundreds of billions of dollars annually, realized through a combination of re-engineering health services and commercial exploitation.

In this article we argue that both assumptions can be challenged. Electronic health care datasets are typically chronically incomplete and unquantifiably inaccurate. This is not always a stumbling block: datasets can still be valuable for a number of purposes, including holding hospitals and other organizations to account, when reasonable accuracy is 'good enough.' Incomplete and innacurate datasets are not, though, good enough to substitute for expensive and time-consuming randomized controlled trials, a key aspiration of health policy makers. In relation to the second assumption, Open Data policies rest on the idea that the state has claims on our personal information. This runs counter to trends in many countries, where there is a strong emphasis on giving individuals access to their own health records, and on the protection of confidential information about diagnosis and treatment. More generally, we suggest that Open Data policies are based on top-down and abstract ideas, which do not take account of the organizational or technological realities on the ground (Margetts, 6 and Hood 2010).

We focus on the prospects for Open Data in the National Health Service (NHS) in England. In common with health systems in many other countries, including most of Europe and the USA, which are also pursuing Open Data policies, it is required to publish much more information about its performance than it has in the past. In the next section we briefly outline historical trends in the production of datasets and the development of the Information Technology (IT) infrastructure within the NHS. Then we set out recent UK Government policies on Open Data in England, which we believe will substantially shape Big Data policies and practices in health care. In order to evaluate the prospects of success we identify three challenges—data management, technology infrastructure, and information governance—and highlight the extent to which each of these is likely

to support or undermine achievement of the Government's objectives. The following section points to ways in which the data landscape is changing, particularly in relation to telehealth, which illustrates the ways in which Open Data is shaping technology policy. Finally, we comment on the nature of Open Data policies, suggesting that they reflect neo-liberal thinking, which emphasizes strong individual property rights, and the need to design economic and legal institutions to reinforce property rights.

Background: Data, Technology and Governance

The NHS has always been a bureaucracy, run from central government, with the Secretary of State for Health being formally responsible for the delivery of health care to the population. As with any bureaucracy, the details do not quite match up to the textbook ideals, with the center issuing commands and expecting them to be implemented. For example, the majority of general practitioners are independent contractors, who happen to earn almost all of their income from seeing NHS patients—and who are universally viewed as being an integral part of the NHS. For the purposes of this article, though, it is reasonable to characterize the NHS as a large bureaucracy, with over 1 million employees. It has clearly defined sub-units, including general practices, hospitals, ambulance services and so on.

Data

For several decades, NHS data and information technology (IT) policies have mirrored the bureaucratic arrangements. Taking data policies first, the top of the bureaucracy—the Department of Health and its predecessor, Health and Social Security—specified data items that had to be collected, and reported upwards. In the early days of the NHS, data collection requirements were relatively limited,

and focused on hospitals, with very little data collected about (independent) general practice and mental health services. In the late 1980s more systematic data collection requirements were introduced, comprising several hundred data items covering all parts of the NHS. The data collection framework was refined and extended in the early 1990s, in response to the introduction of New Public Management (NPM) policies. The new policies sought to separate out purchasers and providers of services, creating an 'internal market' within the bureaucracy, and de-centralizing decision-making. The logic of these changes, at the time, was that purchasers would agree contracts with hospitals and other providers of services. These arrangements required a considerable increase in the volume and quality of data collected—though paradoxically, the Department of Health effectively determined the content of contracts. The competition-promoting elements of the policies were diluted in the mid-1990s, but the structures and data collection requirements were retained (Klein 1998; Webster 2002).

In the 2000s there was a marked increase in the number of datasets that NHS organizations were required to collect. The nature and purpose of the datasets varied greatly, but included data on waiting times for hospital treatment (measured against a main target of 18 weeks from GP referral to clinical action), data from general practitioners to evaluate their performance in relation to their national employment contracts, 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 was centralizing, in two senses of the term: the Department of Health defined what data were to be collected, and the Department and national agencies, rather than local NHS organizations, were key users of the datasets.

We are not aware of any detailed surveys of the uses of these datasets. Our subjective judgment is that there is extensive use of datasets collected for performance management purposes, and highly variable use of datasets for the

review of services for people with particular conditions, or for monitoring quality and safety. We are on firmer ground when we observe that major failures, where hospitals have been found to be providing very poor quality services, were not identified using available data, but were brought to wider attention by whistleblowing staff or patients. We will see later that there are reasons for this state of affairs.

Technology and Governance Failure

It is one of the curiosities of the NHS that data policies have historically been separate from IT policies. Mainframe systems were first used for administrative purposes in the 1960s, mainly for back office functions such as accounting. IT systems were gradually introduced into more and more clinical environments from the 1970s onwards, with early applications in general practice and for managing hospital out-patient appointments. It is worth noting here that, even in the early 1990s, the recording of prescribing and printing of prescriptions was automated, and GPs were able to sell anonymized prescription data to commercial firms. The aggregation and publication of routine data is not new, at least in the UK.

A snapshot taken at the turn of the millennium would have revealed large numbers of systems across most of the NHS, though with lacunae in a few areas, such as community nursing, which were still largely reliant on paper systems. Relatively few of the 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 systems supporting the collation of national clinical datasets, for cancer, heart disease and other conditions, were managed by different organizations, often far removed organizationally from frontline clinicians.

These developments were sometimes encouraged by the Department of Health, but in general they occurred without much central direction. Viewed from above the developments were federal, or bottom-up, in nature. As a result, some systems could be used to collate data required by the center, but in other instances central returns had to be collated manually, taking data from separate systems. An attempt was made to impose a bureaucratic IT model on the NHS in 2002, when the Labour administration of the time launched the NHS National Programme for IT. It marked a move away from NPM policies, replacing them with a more traditional command-and-control model. The Programme was designed to computerize every aspect of NHS services and management.

The story of the Programme has passed into folklore: key systems, notably electronic health record systems, were not delivered, and supplier firms pulled out of the Programme or were fired (Keen 2010; National Audit Office 2011). The less well-publicized fact is, though, that some of the components of a national infrastructure were delivered. In particular, there is now a substantial internal NHS network: summary data from individual systems is passed to a 'Spine,' managed by the Department of Health, and can be accessed there by any organization with the necessary permissions. This has allowed the Department of Health to pursue the creation of 'summary health records,' containing limited data about all individuals living in England (e.g. basic demographic data and recent prescriptions), and to acquire large volumes of data electronically. This mix of automation and centralization has proved to be controversial, but criticism has not led to policy change (Greenhalgh et al. 2010).

From Big Technology to Open Data

In the second half of the last decade the Labour Government changed tack, and emphasized NPM-style policies once more. The current Coalition Government,

comprising Conservatives and Liberal Democrats, continued with these policies after the 2010 general election, extending them to include greater use of non-NHS organizations 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. That is, like earlier administrations, the current Government has been attempting to square the circle of introducing market-oriented thinking into a large bureaucracy. These developments have taken place against a backdrop of financial austerity, and the Government has been looking for opportunities to boost the UK economy.

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 programs are out of favor, centralization 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 a vision for Big Data and Open Data. This includes the planning of services by NHS organizations, commissioning services and research. 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. It envisages the creation of a number of 'portals,' for patients, health professionals, researchers and others to access datasets, though is short on details about their architecture, or indeed who will develop them. And, it emphasizes 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.

Three Challenges

While the quotes in the last section are general in nature, the direction of travel is clear, towards the publication of far more, and more detailed, information than in the past. In order to support the achievement of the Government's objectives, the NHS has to be willing and able to produce and publish detailed datasets. Pharmaceutical and other firms are unlikely to be interested in highly aggregated data, but are interested in data captured in telehealth devices (the Prime Minister

announced a new telehealth initiative in the same speech), in genomics data, and in detailed information about diagnosis and treatment more generally. Is the NHS in a position to deliver? In this section we assess the current status of the NHS under three headings, data management, technology infrastructure, and governance.

Data Management

There are challenges associated with working with all large datasets, whether from the NHS or any other setting. Broadly speaking, these challenges are associated with collection of data, the value of data, the management of previously collected data, and the destruction of (unwanted) data. We consider each briefly in turn. First, much of the data about activity in the NHS, as in other health systems, is collected in busy settings by doctors and other professionals, and by junior administrators. Keying in of data is an expensive and error-prone process, resulting in datasets with many cells that may either be left empty or are incorrect. The automation of data collection is clearly desirable, both for local and Big Data purposes. There have been some advances in this area (e.g., better keypads and user interfaces that are designed to make data entry faster and less error-prone). If data collection can be automated, promising developments in the field of automatic network scanning (Holm et al. 2012) for populating Big Datasets 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. At present, though, the challenges of automation at scale, across the NHS, remain poorly understood from a technological perspective.

A second key challenge concerns the value of collected data. A dataset is, in effect, a model (indeed, there is even a standard for modeling datasets, the Common Warehouse Metamodel, CWM), and a model is always constructed for a purpose (e.g., reporting to the center, commissioning services, detecting unsafe clinical practices). The value 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. In practice, large datasets often serve many purposes, and it can be difficult to understand how to improve or judge their accuracy in relation to any one objective. In the case of NHS data, though, and indeed with datasets in any health care system, there is a tension between two objectives, namely supporting clinicians in the diagnosis and treatment of patients and central reporting. The desire to pursue Open Data policies is, therefore, in tension with the proper desire to design systems that will help doctors and other clinicians.

There is also a significant *freshness* issue associated with Big Datasets: they become stale over time and identifying when they are stale, and when action is required (e.g., disposal of the dataset, refreshing it, auditing it) is always difficult. In principle, the NHS should be generating 'as live' datasets, because any quality and safety problems should be identified and addressed immediately, but we are currently a long way off understanding what models are needed to allow health professionals and managers to do this. Again, this objective will be in tension with Open Data objectives, focused as it is on supporting better care rather than on central reporting.

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 (i.e. not only Structured Query Language, SQL) databases may help to address this. Moreover, the procedures and processes 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 access. There are also significant challenges related to the efficiency of access and management of Big Data: such data are simply difficult to store, and many organizations do not have the facilities to do so. Hence, public, private or hybrid cloud storage services will become increasingly important. In addition, 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? We sense that the answer is: it will depend on the questions being asked. Datasets may well be 'good enough' for some purposes, but not for others, notably in substituting for randomized controlled trials, where internal validity, and hence confidence in the source data, is the prime concern.

Our final challenge is something that we raise as a concern: how and when should we dispose of Big Datasets? 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 and Jenkins 2008), but with Big Data, cross-organization governance concerns will also come into play (Mayer-Schoenberger 2009).

Technology Infrastructure

From an engineering perspective, the advent of Big Data poses exciting new challenges for technology infrastructure. The excitement has a paradoxical quality, because it stems from the fact that we do not currently know how to solve some important problems, and so it is based on promise rather than the ability of computing firms to deliver solutions today. Existing data storage, modeling, querying and analysis paradigms do not directly scale to Big Data problems, 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 become clear here. Technology will have to play 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 modeling, 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, modeled 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 highlighted 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 5 February 2013), 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; for example, privacy protection against statistical inference attacks (i.e., 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 analyze Big Data, powered as it is by new and complex "virtualization" system software, it is likely to contain security loopholes that will take some time to uncover and fix.

Information Governance

There are three distinct information governance challenges in relation to Open Data. The first, and arguably most important, concerns confidentiality. Many health problems are highly personal, and patients need to be confident that their conversations with doctors and other professionals are confidential. By extension, records of conversations and treatments should be confidential. In practice, confidentiality is assured through legislation on data protection and on confidence (i.e., you cannot publish confidential information about me unless there is a clear public interest argument for doing so). Manson and O'Neill (2007) and Brown et al. (2010), among others, stress the complexities of the legal and regulatory frameworks surrounding personal health information.

The crux of the problem is that patient records have two distinct applications, one for treating patients themselves and the other for so-called

secondary uses, including medical research and planning health services (e.g., tracking the latest outbreak of swine flu, or estimating the numbers of people who will need treatment for their diabetes next year). It has proved to be difficult, both in England and other jurisdictions, to strike the right balance between individuals' confidentiality and apparently legitimate secondary uses, even when data are stripped of identifying information. The problem, as already noted, is that modern technologies make it possible to combine data from a wide range of sources, including the publicly available information about many of us on the Internet (Ohm 2009; Zittrain 2008), to the extent that individuals can be identified in datasets, even when considerable efforts have been made to de-identify them. Open Data policies sit on this regulatory fault line. The more detailed information is, the more valuable it is likely to be to third parties, but the more detailed it is the greater also is the likelihood that patients will be identifiable in the data. It is not clear, at present, how the circle of personal data protection and commercially valuable publication can be squared. Indeed, the NHS has little experience of publishing detailed datasets, and new guidance will be needed.

The second governance issue concerns the overall design of NHS governance arrangements. As noted earlier, the NHS was subjected to an NPM-style reform program in the early 1990s, but remains a bureaucracy, financed by central government. It is a moot point whether command-and-control and NPM are really alternatives to one another, or have co-existed uneasily for the last 20 years. The point of making the separation here is to highlight the tension, for Open Data, between the two. The more NHS organizations are required to compete, the more reticent they are likely to be to release information about their performance. Competitors will be interested, for example, in information about a new service delivery model that a hospital is using, and in the cost reductions achieved in delivering those services. Yet the success of Open Data depends on central government being able to demand that individual NHS organizations

publish as much information about their performance as possible. It seems that the success of Open Data will depend on command-and-control trumping competition—running counter to current Coalition Government policies for the NHS.

Third and finally in this section, Open Data enthusiasts can 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) 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 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. Developments in the NHS are consistent with this line of analysis. A combination of government IT policies, the current state of development of the IT infrastructure, and the contested ownership of health care information all contribute to a situation that is far from the rational-technical ideal required to deliver Open Data.

A Changing Landscape

Up to this point we have focused mainly on data held within the NHS. But developments in the use of NHS datasets are taking place in the context of a number of broader trends. One trend is for patients to share information about

treatment options in dedicated services such as PatientsLikeMe (http://www.patientslikeme.com), largely independent of health professionals. A second trend is in the area of genomics. It is now possible to analyze 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. In spite of this, there is funding in the UK for sequencing the genomes of 100,000 patients with cancers or rare diseases, and an explicit intention to exploit the resulting data, both for those patients and also to help to develop new diagnostic tests and treatments (Hawkes 2012). This opens up a new possibility, namely collecting complete datasets—individuals' DNA sequences—from selected individuals. This might circumvent some of the problems set out in earlier sections, though it would not of itself solve others.

Similar thinking informs policies in the area known as telehealth, where the computing and medical devices industries are gradually integrating with one another. Free-standing medical devices that monitor blood sugar levels, oxygen and other variables are increasingly being promoted for use in peoples' homes, linked to computer networks, such that individuals' health status can be monitored remotely. The UK Labour Government funded a randomized controlled trial of telehealth, which started in 2008. The Coalition Government anticipated positive results, and in late 2011 announced that it would support a program to provide funding for up to three million people with chronic health problems to use telehealth devices and services (3millionlives.co.uk). As with genome data, it was made clear that the data from these devices were deemed to be commercially valuable. In the event, the findings of the trial were mixed at best (Steventon et al. 2012)—but it is clear that the initiative will go ahead. And, again in common with

genome data, it is difficult to escape the conclusion that the desire to acquire new, complete datasets from a commercially important population has trumped a more sober assessment of the cost-effectiveness of a technology.

New Public Management and Beyond

What do statements about Open Data policies, and the challenges that they throw up, tell us about the nature of the policies? A distinction has already been made between NPM and command-and-control policies. Both policies—in essence different sets of organizing ideas about NHS structures and processes—have been emphasized at different times, with NPM policies currently dominating. Open Data policies contain elements of both sets of ideas, with the emphasis depending on the vantage point you choose. Viewed from central government, Open Data policies are centralizing, consistent with command-and-control thinking. Success will depend on the center being able to address the challenges set out in the last section. If one takes a broader perspective, though, that encompasses the whole health care sector, then Open Data policies appear to have more in common with NPM. Open Data looks like a new form of public–private partnership, with central government brokering the arrangement, with private interests able to influence the content and design of datasets.

We sense, though, that Open Data policies are broader in scope than NPM and command-and-control ideas. They are concerned with the success of UK businesses, rather than with intra- or inter-organizational governance challenges. We are looking, then, for a framework with appropriate breadth. We suggest that Open Data policies are consistent with a particular strain of neo-liberal thought. Neo-liberal theory emphasizes strong individual property rights, freely functioning markets and free trade, and strong legal institutions to underpin the property rights and markets. Harvey's (2005, 64-86) analysis 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 resolutions interest us here. First, if tensions arise between individuals and firms, for example over property rights, those tensions tend to be resolved in favor 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 favorable to those firms.

Open Data policies exhibit features of this 'practical neo-liberalism.' The quotes from the UK's Prime Minister and Chancellor of the Exchequer highlight the importance of property rights over personal information, and attempt a resolution that is favorable to private firms. They also emphasize cooperation 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 five interests, namely the state, private firms, citizens/patients, doctors and other health professionals, and a broader diaspora of users including health charities and journalists. Open Data reinforces relationships between state and private sector actors, and seeks to do so by weakening the positions of both citizens/patients and professionals.

Conclusions

At the beginning of this article we identified two assumptions, one concerning the infrastructure for collecting and managing data, and the other that it would be

possible for third parties to access detailed datasets. What can we say about the prospects for Open Data policies in the NHS? In relation to the first assumption, the long history of extensive centralized data collection, and the substantial number of new regional and national datasets created in the last decade or so, help to make the point that the NHS has always collected and managed Big Data, and is able to coordinate the collection of new datasets from large numbers of organizations when required to do so. At the same time, and as our comments have emphasized, this history allows us to identify substantive data, technology and governance challenges. Some of these are bound to be solved over time, but we may have to live with others, notably with tensions over beliefs about the ownership of information.

In relation to the second assumption, we have argued that success with Open Data policies depends on political decisions to make datasets available to third parties. This leads us to the equation in the title of the article. Big Data and Open Data are both presented, in the wider media, as being essentially technological ventures. While this seems to be a reasonable description of Big Data activities, with their established history inside the NHS, Open Data policies are inherently political. In order for them to be successful they will require the Government to assert property rights over large datasets, in a way that it has not before, and to require NHS organizations to collect and hand over data requested by non-NHS organizations. Indeed, Open Data policies could over time provide a route for non-NHS organizations to influence data collection and management within the NHS. The Prime Minister's support for collecting data from 3 million people using telehealth devices may be an early example of this type of development. The prospects for Open Data depend, therefore, on the Government controlling NHS data management—Big Data plus politics produces Open Data.

This brings us to our last point. The success or failure of Open Data in the NHS may turn on the question of trust in institutions. Trust will have to be

negotiated between the parties identified in the last section—the state, private firms, citizens/patients, doctors and other professionals, and the broad set of those with interests in the conduct and performance of the NHS. An optimist might argue that the NHS represents a successful, long-standing political settlement between these same interests: we mostly get good treatment, professionals have security of employment, firms are able to sell their goods to the NHS, and so on. Open Data implies the need for a similar settlement, focusing on information rather than on the use of our income taxes for diagnosis and treatment. The difference is that we trust health professionals, and particularly our general practitioners, but Open Data requires us to trust the government of the day and private firms. This trust may not be easily won.

References

Anderson S, G Hardstone, R Procter, R Williams. 2008. "Down in the (Data)base(ment): Supporting Configuration in Organizational Information Systems." In *Resources, Co-Evolution and Artifacts*, 221–253. London: Springer.

Brown I, L Brown, D Korff. 2010. Using Patient Data for Research Without Consent. *Law, Innovation and Technology* 2(2) 219-258.

Cabinet Office. 2011. *Cloud Computing Strategy*. http://bit.ly/vFSQ28

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

Department of Health. 2012. *The Power of Information*. London, Department of Health.

Greenhalgh T, K Stramer, T Bratan, E Byrne, J Russell, H Potts. Adoption and non-adoption of a shared electronic summary record in England: a mixed method case study. BMJ 2010;340:c3111.

Hawkes N. 2012. Cameron announces £100m for "unlocking the power of DNA data". BMJ *BMJ* 2012;345:e8413

Holm H, M Buschle, R Lagerström, M 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.
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.
Klein R. 1998. Why Britain is Reorganizing its National Health Service—Yet Again. Health Affairs 17, 111-25.
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.
Mayer-Schonberger V. 2009, Delete. Princeton NJ, Princeton University Press.
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.
Star S. 1999. "The Ethnography of Infrastructure." *American Behavioural Scientist* 43 (3): 377–391.
Ohm P. 2010. Broken Promises of Privacy. Responding to the Surprising Failure of Anonymisation. UCLA Law Review 57, 1701.
Steventon A. et al. for the Whole System Demonstrator Evaluation Team. 2012. Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial. *BMJ* 2012;344:e3874
Topol E. 2012. The Creative Destruction of Medicine: How the Digital Revolution Will Deliver Better Health Care. Basic.
Webster C. 2002. The National Health Service: A Political History. Oxford, Oxford University Press.
Zittrain J. 2008. The Future of the Internet. London, Penguin.