Information Polity 26 (2021) 57–69
DOI 10.3233/IP-200264
IOS Press

Machine learning, materiality and governance: A health and social care case study

Justin Keen[a,*], Roy Ruddle[b], Jan Palczewski[c], Georgios Aivaliotis[c], Anna Palczewska[d], Christopher Megone[e] and Kevin Macnish[f]
[a]*Leeds Institute of Health Sciences, University of Leeds, Leeds, England*
[b]*School of Computing, University of Leeds, Leeds, England*
[c]*School of Mathematics, University of Leeds, Leeds, England*
[d]*School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, England*
[e]*School of Philosophy, University of Leeds, Leeds, England*
[f]*Centre for Ethics and Technology, University of Twente, Enschede, Netherlands*

Abstract. There is a widespread belief that machine learning tools can be used to improve decision-making in health and social care. At the same time, there are concerns that they pose threats to privacy and confidentiality. Policy makers therefore need to develop governance arrangements that balance benefits and risks associated with the new tools. This article traces the history of developments of information infrastructures for secondary uses of personal datasets, including routine reporting of activity and service planning, in health and social care. The developments provide broad context for a study of the governance implications of new tools for the analysis of health and social care datasets. We find that machine learning tools can increase the capacity to make inferences about the people represented in datasets, although the potential is limited by the poor quality of routine data, and the methods and results are difficult to explain to other stakeholders. We argue that current local governance arrangements are piecemeal, but at the same time reinforce centralisation of the capacity to make inferences about individuals and populations. They do not provide adequate oversight, or accountability to the patients and clients represented in datasets.

Keywords: Machine learning, governance, accountability, information infrastructure, health care, social care

Key points for practitioners

- Machine learning tools can increase our capacity to make inferences from personal datasets.
- The poor quality of routine datasets limits the potential of these tools.
- Current governance arrangements are piecemeal.
- The use of machine learning tools is likely to reinforce already-centralized information infrastructures.
- The centralized infrastructures tend to disadvantage key stakeholders, including the people who are represented in datasets.

1. Introduction

A range of academics, policy makers and other commentators believe that a combination of machine learning tools and access to large personal datasets will produce novel insights in many domains, including

*Corresponding author: Justin Keen, Leeds Institute of Health Sciences, University of Leeds, Level 10, Worsley Building, Clarendon Way, Leeds LS2 9JT, England. Tel.: +44 113 343 0831; E-mail: j.keen@leeds.ac.uk.

1570-1255/$35.00 © 2021 – IOS Press. All rights reserved.

health and social care (AHSN Network 2020, Future Advocacy 2018, O'Neil, 2016). The claims, often implied rather than explicitly stated, are that these insights cannot be produced using established statistical and other methods, or that the new tools will out-perform established methods in some way. Considerable sums of money are being invested by governments and private interests, in many countries, to exploit the opportunities. At the same time, concerns have been expressed that analyses are being undertaken without the consent of the people represented in datasets, and privacy and confidentiality are being undermined. Analyses of personal datasets might reveal sensitive information about individuals, which in the case of health and social care might concern a recent sexually transmitted disease or the details of someone who has sought shelter from an abusive partner. Critics worry about the long-term consequences for autonomy and for trust in institutions if personal data are perceived to be misused or abused (Ezrachi & Stucke, 2016). It is not currently clear how, or even whether, the claims and concerns can be reconciled, such that the benefits and risks associated with new tools and datasets can be balanced with one another.

Much of the academic debate about personal information to date has focused on consent, privacy, confidentiality and trust, and philosophers and socio-legal scholars have provided helpful clarifications of these concepts (Kennedy, 1994; O'Neill, 2002). Guidelines for the design of regulations, drawing on these concepts, have also been published (Mittelstadt & Floridi, 2016; Fernow et al., 2019). Less attention has, though, been paid to the granular details of current and possible future governance arrangements, and in particular to the ways in which the material properties of data and technologies might influence those arrangements. This article investigates governance arrangements for the use of new machine learning tools, designed to interrogate personal datasets in health and social care. We argue that the material properties of datasets and software tools both enable and constrain the ways in which they are used. Drawing on a historical institutional approach, we observe that information infrastructures have developed over more than twenty years in piecemeal fashion, with technology artefacts and working practices shaping one another (Hanseth & Ciborra, 2007; Hyysalo, 2015). Governance arrangements, similarly, have developed piecemeal over time, and a historical perspective allows us to identify their strengths and weaknesses. For example, we argue that there has been a tendency favour 'vertical' data flows from ground level services to central government agencies, and that this has been at the expense of accountability for the use of datasets to other stakeholders, including representatives of patients and clients.

We then discuss research undertaken in the Quanticode machine learning and visualisation project (Adnan & Ruddle, 2018; Palczewska et al., 2017). The project developed tools that can be used by partners in health and social care organisations. One strand of research ran alongside tool development, and focused on the principles and practicalities of creating what we termed 'ethically aware tools'. We found that they enhanced the capacity to make inferences about individuals and groups of people. Conversely, the poor quality of routine datasets limited (but did not eliminate) the potential of the tools to make inferences. And, the methods used and outputs of analyses were difficult to explain, even to technically knowledgeable colleagues. Putting the historical arguments and Quanticode findings together, we conclude that the centralising tendencies of information infrastructures in health and social care and the advent of machine learning are mutually reinforcing. This has tended to exclude the involvement of, or accountability to, other stakeholders. We further conclude that current governance arrangements are piecemeal, and are not appropriate for the task of regulating the new combinations of tools and datasets, particularly in the absence of voices for patients and clients.

2. Methods

In order to understand the material effects of personal data and associated information technologies, we developed an account using the Biography of Artefacts approach (Pollock & Williams, 2009). The

approach is based on the observation that modern information technology (IT) systems are not just implemented and then periodically upgraded. Rather, new functions are added to the initial system over time, and new systems are implemented and linked to the initial system, often in piecemeal fashion. An organisation might, for example, start with a web site and servers for storing work documents, and then implement a payroll system, a human resources system and specialist software for one of the teams over a period of years. Many members of staff will need to access two or more of these systems from their desktops, requiring informatics teams to consider system integration, user interfaces and other hardware and software issues. These developments result in digital infrastructures that become deeply embedded in the day-to-day work of organisations, composed of partially integrated elements. The consequence is that, if we want to understand any particular set of systems, and understand why it looks the way it does today, we need to understand its history.

An initial Biography of developments in information infrastructures for managing administrative data in health care has been published elsewhere (Keen et al., 2018). It shows that there is a long history of hospitals and other organisations submitting data to higher levels in bureaucracies (or to insurers), that the scope of data submitted has increased over time, and that there has been a re-purposing of the use of that data, notably from supporting administration to performance management. In England formal central data collections started in the 1980's with a limited set of hospital activity and finance data. In the intervening period the volume of data submitted by hospitals has increased substantially, and family physicians and other organisations now routinely submit datasets as well. Here we summarise key points, and extend the account to cover developments in the governance of secondary uses of datasets over the last 20 years, and the emergence of new tools and datasets, notably in genomics and the domain of machine learning.

The material properties of machine learning tools and datasets, and their implications for governance, were investigated at a more granular level. We developed machine learning and visualisation tools in the Quanticode project. The tools were designed to enable end users in health care, social care and other organisations to analyse quantitative and coded longitudinal data. Methods for mining large datasets were developed, as were interfaces designed to allow managers and analysts to interrogate their own data in new ways, using the new analytical tools. Key technical results are reported elsewhere (Adnan & Ruddle, 2018; Adnan et al., 2019; Palczewska et al., 2017; Ruddle & Hall, 2019). The strand of research reported here is based on within-team discussion and reflection during the project. The team included mathematicians, computer scientists and philosophers. We discussed the properties of the new tools, their governance implications and the extent to which our experiences reflected those reported in wider literatures on machine learning and governance.

3. The development of information infrastructures

In health and social care services there have historically been two main – and largely separate – information infrastructures. The first, familiar to any patient or client, is the personal record. The development of health records, in particular, has been extensively studied (Boonstra et al., 2014; Otero-Varela et al., 2019). Starting with the recording of basic transactions – admissions, discharges and deaths – in hospital ledgers in the eighteenth and nineteenth centuries, they developed over time into separate paper records for each patient. Today records are typically still held on paper in many hospitals, but increasingly also on servers, with the latter including test results, x-ray and other images, and the results of nursing and other assessments.

Historically, people were typically treated in one place, either in the community or in an institution, and most had just one record. In the last forty or so years, though, it has become usual for people to live

at home with their illnesses, and to receive treatment and care from a number of different professionals, often working for a number of different organisations, in parallel. This is the case for the significant minority of people who have one or more long-term conditions including cancer, heart disease, mental health problems, diabetes or neurological conditions such as Parkinson's disease. The result has been a proliferation of records, so that any one patient may have records held by her family physician (general practitioner), local hospital, community nursing service, social service and others. As we note below, in the last decade or so significant efforts have been made to provide professionals with access to one anothers' records.

The second, separate type of infrastructure was developed initially to support the administration of services. The history stretches back to the 1960's, when early mainframe systems were used for 'back office' functions such as managing staff payrolls, and for the aggregation of basic data, particularly about hospital activity. There were significant developments in the 1970's, particularly in hospitals, including the introduction of patient administration systems - to manage appointments and admissions - and separate systems for recording activity in operating theatres and in other departments. These were data processing systems, designed to capture and store discrete data items, perform relatively limited computational operations and produce summarised outputs.

Developments in hospital systems continued in the 1980's, notably with the emergence of financial management systems, designed to link activity and costs (Packwood et al., 1991). Indeed, finance managers were prepared to fund investments in departmental systems, in part because they understood their value. This was also the decade in which the first general practice (family physician) computer systems were developed (it should be stressed that the extent of automation varied greatly between countries, right into the 2000's). These functioned both as local patient records systems, available to doctors on their desks, and as sources of data for the administration of payments to general practices, by enabling the generation of datasets that were then submitted to a payment authority. The latter was part of a broader development, of data submissions from general practices, hospitals and community services, which were used for payment purposes and/or to construct basic performance indicators. Initially many of these submissions were on paper, but over time they were automated – submitted on disks or tapes – and eventually transferred computer-to-computer via data networks. As Bowker and Star (1999) have pointed out, this was also a period when inherently fuzzy phenomena such as clinical diagnoses began to be standardised, making it easier to represent them within data processing systems.

If we had taken a snapshot of the digital landscape in health and social care at the turn of the century, we would have observed automation of many health and social care services. Most could still be accurately described as data processing systems, existing largely in islands of automation (to use an old-fashioned phrase). A supplier market had developed, populated by a range of large and small hardware and software suppliers. There were still gaps, particularly in community services, where health and social care professionals continued to rely on paper systems. There were integrated clinical information systems in some hospitals, notably in larger hospitals in the USA and (a few) European and east Asian cities, but these were the exception rather than the rule. The volume of data submitted to regulators and insurers has continued to increase. In England, for example, a number of new initiatives were introduced in the 2000's, including disease registers for diabetes and other clinical conditions, and reporting of safety incidents (Keen et al., 2018).

Today, the majority of acute hospitals in developed economies are extensively automated. They have patient administration systems, systems for ordering and reporting test results, and there are dedicated clinical systems to support treatment and care in wards and departments. It should be noted here that paper is still ubiquitous, with most hospitals retaining paper patient records alongside the IT infrastructure.

IT systems are now indispensable for back office functions, including finance and workforce planning. Outside hospitals, automation of general practice is universal in many countries. It is now common, but still not universally the case, for community health and social services providers to have laptops or other portable devices, that allow them to either carry personal records with them or to access patients' and clients' records remotely. The range of data items captured is, though, still influenced by what one might term 'data processing thinking'. That is, whether a health system is tax-financed or insurance-based, administrators and insurers specify a substantial proportion of all data items captured, reflecting their information needs rather than those of doctors and other clinicians on the ground. Moreover, those data are typically activity counts, such as operations performed or older people who have entered a nursing home, reflecting a style of thinking based on the centre's desire to account for activities. This in turn influences the designs of systems and the data captured in those systems used for direct care by health and care professionals.

As noted above many people, notably people with chronic problems, are supported by a number of services at any one time. There is therefore a need to co-ordinate services across traditional bureaucratic boundaries, such as those between primary and hospital care, and health and social care (O'Hara et al., 2019). This has led to efforts to develop a third class of information infrastructure – over and above patient records and administrative infrastructures – to support the co-ordination of treatment and care across organisational and professional boundaries (Bates, 2015; Walker et al., 2005). There have been developments in inter-operability in the last decade. In many countries GPs are able to 'view' their patients' hospital records, including recent pathology results, remotely; hospital staff can, similarly, view GPs' records. Taking England as an example, it is now usual for health and care professionals to be able to view records held by other organisations remotely. In general this facility is based on functionally interoperable networks, where access is to records in the format used by the host organization (Keen et al., 2019). It seems reasonable to observe that these infrastructures are still in an early stage of technological development in most localities.

3.1. Governance arrangements

Routine data about public services have a reflexive role in governance. On the one hand they provide evidence about the performance of services, and today many hundreds or even thousands of data items are routinely submitted to regulatory bodies or insurers by hospitals and other organisations. On the other hand, datasets are aggregations of personal data, and accordingly their use has itself to be governed, to promote effective use and to avoid misuse or abuse. A number of trends, including the availability of increased processing power and storage capacity, and the increasing use of personal data in online transactions, have led governments to introduce data protection legislation and associated regulations since the 1980's. These have culminated in the European Union's General Data Protection Regulation (European Union, 2016), which is increasingly being used as a template for regulation by other countries around the world. (The Regulation is not perfect, not least in failing to provide a useful working definition of 'personal' data, a problem highlighted by Manson and O'Neill (2002) many years ago.)

Further oversight of the release of health datasets has been introduced, in the form of research ethics approval arrangements (in England these only became mandatory in the 2000's). Oversight of the release of social care datasets is still less well developed, and is largely the responsibility of individual local authorities. This has provided some assurance that research projects have ethical objectives, and that personal data obtained in the course of a project will be handled ethically. Ethical approvals emphasise the importance of obtaining patients' and clients' consent to the use of data about them, or require stringent

constraints on the use of datasets when consent cannot be obtained (because, for example, many thousands of people are represented in routine datasets). It is also usual to have to obtain formal approval for the release of datasets from the bodies that hold them: firms, universities and others have to demonstrate that they can hold datasets securely and will use them for legal purposes.

The result, typically, is the release of a dataset that has been 'anonymised', which in practice means that obvious personal identifiers such as names, addresses and dates of birth have been removed. The reasoning here is that, taken together with appropriate security measures – essentially, making it very difficult to access a dataset without the appropriate permissions - the risk of misuse of a dataset is greatly reduced. If datasets are to be useful, though, they need to retain information about individual patients and clients, so that it is possible to undertake analyses that account for variables such as age (month of birth may be retained) and gender, and to find useful patterns of diagnoses, treatments and outcomes. The main risk is that patients or clients might nevertheless be identified, based on this still-rich information, opening up the in-principle possibility of misuse.

Just a few years ago, data releases were not deemed to be a public policy issue. NHS bodies had been releasing datasets to universities and private firms since the 1980's. Tanner (2017), similarly, observes that citizens in the USA are often unaware that personal data are routinely sold to third parties. The situation changed, in England, with widely publicised cases that suggested either that data protection and other laws provided insufficient protections for personal data (in the view of journalists and the wider public), or that laws were being broken. They raised questions about who decides on the uses of personal datasets (Davis, 2017; Kaplan, 2016). An example of the former is an attempt by a government agency in England to link personal data from family physicians to hospital data, in a policy called care.data (Health and Social Care Information Centre 2014). This linkage was perfectly legal, but was attempted without any effort to explain why it was being done, and when questioned ministers and civil servants were unable to provide assurances about the release of linked datasets to third parties. A decision was made not to obtain the explicit consent of citizens to link data (even though family physicians had agreed in principle to obtain consents). The policy was abandoned in the face of considerable public opposition, with a small but politically significant proportion of people (over one million, in a population of some 55 million) opting out of the policy.

An example of laws being broken is the release of large NHS datasets, of around 1.6 million patients, by the Royal Free NHS hospital to Google DeepMind. The avowed purpose of the release was to develop a software tool to support the diagnosis of kidney infection in hospital patients. But, data were released to DeepMind without *any* formal approvals being sought or obtained (Powles & Hodson, 2017). The relevant regulator, the Information Commissioner, found that the hospital had breached data protection law.

It is worth putting these high profile events in context. There is evidence from the UK, based on reviews of large numbers of data releases, that suggests that most data breaches are accidental, and only a small minority egregious (Laurie et al., 2015). The examples given here are the exceptions to a generally more positive rule. The exceptions do, though, provide important context for our work, reported in the following sections. They have highlighted that oversight is piecemeal, with different bodies responsible for different aspects of oversight, not actively co-ordinating with one another. Moreover, in the last few years new datasets and analytical methods have emerged, which have further sharpened debates about the design of regulatory frameworks, raising questions about the extent to which the current piecemeal framework is still appropriate. One development, beyond the scope of this article, involves the capture of data by social media companies, either through users providing it by typing it in, or through capture via devices (such as Fitbits and Apple Watches) that measure or estimate movement, heart rate and other variables.

Similarly, genomic data have long been used to support the diagnosis and treatment for individual patients with genetic disorders, but have generally only been aggregated on an *ad hoc* basis in research projects. Infrastructures for handling the genomic data of many thousands of people, including biobanks, have developed rapidly in the last few years (Genomics England, 2020; Bycroft et al., 2018). A third example, to which we now turn, is the emergence of machine learning tools.

4. Machine learning tools: Materiality and governance

Machine learning refers to mathematical models that are designed, initially, to identify patterns of interest in a dataset: they are 'trained' on the dataset. Some of the early models have been designed to perform analyses that are not possible using conventional statistical methods. They have, for example, been trained on hundreds of digital images of retinas of who already have a diagnosis of a particular condition, so that commonly occurring patterns can be associated with the diagnosis. The resulting tools have been used to support diagnosis, by identifying similar patterns in images from patients who have not yet been diagnosed. There is evidence that tools can accurately identify some conditions, although at present it is still best to rely on a combination of tools and humans (Badar et al., 2020).

Some of the excitement about machine learning concerns the prospects for finding novel patterns in – and being able to make inferences from - personal datasets (Selbst et al., 2019; Malik, 2020). In practice this means using the routine datasets produced in the administrative information infrastructures described above. This will, advocates argue, enable mathematicians either to identify relationships in datasets, or make (cognitive) inferences from them, for example by making inferences about our health and wellbeing from data about our patterns of use of services.

Our experiences of developing machine learning tools and visualisations are set out under three headings, namely acquiring datasets, developing tools, and implications for governance arrangements. The Quanticode project involved working directly with partners who held health and social care datasets, and the findings reflect our experiences of working with them. The accounts below include our interpretations of our experiences, where they were part and parcel of this strand of the research project.

4.1. Acquiring datasets

Acquiring health and social care datasets involved negotiation with a patchwork of organisations including ethics committees, organisations holding datasets and accreditation bodies (eg for ISO27001 on data management), as well as our project partners. The state of affairs has been noted in some academic literatures, including health services research and socio-legal studies, but appears not to be widely known in academic computer science and mathematics circles (Fazlioglu, 2019). Initial discussions with the various stakeholders indicated that it would be possible to acquire single anonymised datasets, that did not require any linkage before or after release to us, within the timescale of the project. Requests that involved linkage of datasets held by different organisations, in contrast, were likely to take longer to approve, or might not be successful. This was partly because it was not clear, to those organisations, who would be responsible for linkage. As we note below, there were arguments for linking NHS (National Health Service) and social care datasets, held by different organisations. While there is UK legislation that covers the secondary uses of all datasets, there is also health-specific legislation: the Health Act 2001 and subsequent legislation allows the NHS to retain data for longer than other organisations, partly in order to allow it to maintain longitudinal datasets for research into cancer, heart disease and other

conditions. In practice, though, neither general nor NHS-specific legislation provides a basis for linking NHS and non-NHS datasets.

As a result, there was no body that felt that it was in a position to take responsibility for the linkage of health and social care datasets. We made the pragmatic decision to request only single, anonymised, datasets and hence only request release of a dataset from a single organisation to our university. Negotiations nevertheless proved to be time-consuming, and involved extensive form-filling and post-form checking (and re-checking) of details. The forms, over and above the original research proposal, covered ethical approval by our university, data sharing agreements and sometimes also data access applications. These were backed up by paperwork, provided to external organisations, describing the security arrangements for data storage and access in the university. Once we had worked our way through the approvals some datasets arrived within weeks, but one took over a year from the time that release had – as we understood it – been approved.

There were other issues which contributed to the significant time costs. Partners did not initially understand the governance of research projects, and in particular the need to draft data sharing agreements and other paperwork – such as ethics applications – carefully. Some versions of agreements referred only to staff – thus excluding students – and assumed that we would only need datasets for the duration of projects (and not for a period afterwards, for example to cover time for revision of research articles following peer review).

To set against this there were also positive – if also time-consuming – aspects of these processes. We organised meetings with partners to determine project objectives (which went into more detail than had been possible in the research proposal). This included the research team writing the 'benefits of research' section, then asking partners to re-draft it in their own language, with us then checking the wording before submission. This helped to ensure that the potential benefits were both 'real', reflecting partners' expectations, and realistic, reflecting our knowledge of the mathematics and visualisation that would be involved.

These experiences led us to make two observations. First, oversight was piecemeal. As far as we were aware, there was no co-ordination between the bodies that approved data release and use. Second, governance was focused on the release of datasets rather than on their subsequent storage and use. (Only one body, NHS Digital, responsible for NHS national level datasets, undertook post-release audits.). Governance post-release relied on the research team to voluntarily report any issues, such as an accidental data breach. Our interpretation was that individual organisations were managing their own risks, in significant part by requiring us to assume them.

We also observed that the health and social care data domain is contested. We came across genuine differences of opinion, particularly about consent, privacy and confidentiality, which manifested in practice as different evaluations of the riskiness of holding, or performing analyses on, personal datasets. At one end of a spectrum were the risk-averse approaches to releasing datasets noted above, but we also encountered the view that risks were over-stated. We interpreted the latter as local, and comparatively benign, versions of the views held by those who released data in the DeepMind incident noted above (Powles & Hodson, 2017). Putting this in broader terms, if you don't believe that you have any automatic rights over a dataset, and that other actors – such as people represented in the datasets – do, you may feel that you need to go out of your way to ensure that stakeholders are aware of, and support, your proposed analyses. If, on the other hand, you believe that you have property rights over a given dataset, you may not believe that you need to obtain anyone's consent to analyse it. This belief appears to have driven high profile health care dataset releases in England (Powles & Hodson 2017; Health & Social Care Information Centre 2014). We could not identify a way of resolving this value-based tension.

4.2. Tool development

The possibilities and problems associated with machine learning and visualisation became apparent in practice. On the plus side it was possible to develop new tools, and to generate novel insights that were valued by partners (Ruddle & Hall, 2019). We showed that it was possible to improve the accuracy of prediction of need for intensive social care services significantly (Palczweska & Palczewski, 2019). Previous studies (e.g. Bardsley et al., 2011) had used linked health and social care datasets, but their conventional statistical methods and data quality problems had resulted in poor predictive power. The true positive rate of prediction of people who would need intensive social care was 80%, a more than four-fold improvement. In addition, we developed a framework for mining temporal patterns when datasets were known to have inconsistencies and errors in recording time-stamps (Palczewska et al., 2017). We also implemented an optimised algorithm for mining of frequent temporal patterns with errors in time-stamps and temporal constraints (Titarenko et al., 2019).

A key aim of the project was to automate the analysis of complex datasets. Model development, including the development of machine learning models, typically comprises four processes – prepare the datasets, design the model, train the model, run the model and interpret the outputs (Sacha et al., 2017). Taking data preparation first, it includes the data that are needed to train a model and the data about which predictions are made when a model is subsequently run. One of our interests was in missingness in datasets, so for us data preparation of both health and social care datasets was an intrinsic element of the research. In the course of preparation we were able to develop novel visualization designs, which were instrumental in revealing previously undetected data quality issues across very high-dimensional data (116 variables) in multiple data extracts (Ruddle & Hall, 2019). We also developed a visualization tool, which allowed users both to detect and to explain rarely occurring but important problems with missing values. Stakeholders – the dataset providers, in this case a health care organisation – were able to feed back lessons to frontline staff, with the aim of improving future data quality.

Less positively, issues included large numbers of records in a health care dataset (over 10% of the total) that could not be used, because there were no primary diagnoses, inconsistent encryption and widespread use of invalid characters. Tens of hours of input from a domain expert were essential for reducing the complexity of some datasets so that modelling was feasible. This reduction had twofold aim. First, the volume of data in administrative systems is so large and the number of data items so relatively small that no modeling approach could abstract useful information from the data. Second, almost identical information (from a modelling rather than administrative perspective) could be coded in different ways depending on circumstances such as the urgency of intervention, legal or reporting framework (changing over the years that data had been collected) or even on the person inputting the data. Weeks of researcher time were spent identifying, understanding and transforming (often undocumented) special values in datasets (for example 01/01/1800 for missing dates) so that outlying values were interpreted correctly and did not distort model training and outputs.

Turning to model design, we concluded that here is no single ‘best’ method and choices made by analysts cannot be easily communicated to stakeholders, yet the assumptions (and hence outputs) of each method may be fundamentally different. For example, some methods treat missing data properly (as missing values) but others require the data to be given special values, which are imputed or records discarded (e.g. random forest methods). Further, limits to the granularity of a variable due to software constraints (e.g. there is a 53 category limit in the R randomForest package) or due to insufficient amount of data, meant that variables with more categories needed to be treated in particular ways (e.g. combining rare categories) or discarded. Validation of models is another contested area where communication of

results is difficult and objectives of researchers and stakeholders are not necessarily aligned. In our work with social care partners, our validation procedures centered around accuracy of prediction of a breakdown of independent living. We identified individuals with a very low and a very high probability of this happening and designated the remaining ones as members of an 'uncertain' class. Our partners appreciated the results, finding the 'uncertain' group particularly helpful, as the group narrowed down the list of clients for whom interventions might extend the time that they could live independently.

4.3. Reflections

We noted at the start of the article that the project included a workstream focused on the ethics and governance implications of our experiences with tool development. Taking our cue from Dourish's arguments (Dourish, 2016; Dourish & Cruz, 2018), we make three observations based on discussions in that strand of work. First, both the possibilities and problems associated with machine learning and visualisation became apparent in practice. On the plus side it was possible to develop new tools, and to generate novel insights that were valued by partners, reflected in follow-on projects with some of the partners. At the same time, the characteristics of the datasets constrained our analyses. Some of these constraints, such as missingness, were anticipated. Others stemmed from constraints on linking datasets under current regulations. For example, we worked with a local authority on the likelihood of an older person living at home needing to move into a nursing home. In practice, many people had not been judged to be at risk at the time they moved into a nursing home. We found that potentially relevant data were not available. For example, people who were paying for their own nursing home care did not appear in local authority-funded lists, so that datasets of people living in nursing homes were incomplete. Information about informal carers, typically a spouse or other family member, was typically either limited or not provided. This is important because events in a carer's life, such as them breaking a leg and being unable to care, would be available in family physician or hospital records. The event might precipitate a cared-for person needing a nursing home place, but could simply not be predicted from the available data.

Second, methods and outputs of machine learning are inscrutable, even with the additional of a range of visualisations of the outputs. The General Data Protection Regulation (GDPR, EU 2016) enshrines a right to explanation about data held about you, and the analyses being performed. But the nature of machine learning methods and outputs meant that it was difficult to explain them, even to technically knowledgeable partners (although it should be noted that we had confidence in the results themselves). It was, in turn, difficult for partners to know how much trust to place in the results: they found some results helpful, as noted above, but that was not the same thing as having confidence that they could design real-world interventions on the basis of those results. We note that the problem of explaining methods and results would be multiplied for explaining methods to the people represented in the datasets used.

Third, and looking ahead, our experiences suggest that machine learning will increase the need to develop regulatory frameworks that balance the risks and benefits associated with the analysis of personal datasets. On the one hand, we are persuaded that improvements in machine learning tools will increase the capacity to make inferences over time. The main constraint at present is data quality in routine datasets, but there is evidence that data linkage can improve accuracy and completeness significantly (Pfoh et al., 2014). If regulation becomes more permissive, and it is easier to link personal datasets, then more inferences will be made. To continue with our earlier example, if we could have linked datasets with information on both informal carers and cared-for people in a locality, we might have been able to make more accurate and/or more timely inferences about them. As also argued earlier, this places a greater onus on national and local governance arrangements.

5. Discussion

This article has focused on the relationship between the material properties of information infrastructures, and of machine learning tools and routine personal datasets, and their governance implications. It seeks to fill a gap in the literature on the uses of health and social care datasets, which tends to discuss technologies in broad terms, and focuses on ethical issues more than on governance (Dourish, 2016; Mittelstadt & Floridi, 2016). Tracking developments in the data and technology elements of information infrastructures we find that they have long been, and are still, based on an old-fashioned data processing model. Datasets are still typically organised functionally, representing long-standing organisational boundaries rather than – for example – being organised by patient or client (Keen et al., 2018). The trajectory of developments emphasises that secondary data flows are 'vertical'. Data about treatment and care are captured at points of service delivery – by community practitioners, hospital staff and others – and then conveyed to local informatics staff, who curate it for local managers, and beyond to regulators or insurers.

The net result is centralisation, and only government agencies, regulators (or insurers in other countries) have extensive access to datasets. They are likely to have the resources to fund the development of new tools, as well, and have incentives to use them, perhaps most obviously to monitor the performance of public services. They largely determine what data, and in what formats, are captured and submitted by service providers. This might be characterised as a cybernetic model of governance where – following a cybernetic analogy - data are supplied by organisations (at the ends of chains of nerves) to a central 'brain', which processes and analyses the data, arrives at judgements and issues instructions back down to the organisations (Richardson, 1991). This leads to the first of two conclusions, namely that the new tools and datasets tend to reinforce the centralization of authority, with little countervailing authority invested in other stakeholders, who might hold central bodies to account.

On new tools and datasets, our findings resonate with reports from other research groups (Shah et al., 2019, Wiens et al., 2019). We noted above that governance arrangements were piecemeal, and that the various approval bodies were inclined to assign risks to research teams. As a result, the arrangements relied on the willingness and ability of research teams to police themselves. Moreover, there was no means of accounting for our working practices to key stakeholders, including representatives of the people described in the datasets, e.g older people living in the local authority area. Our second conclusion, then, is that current governance arrangements are not designed to address the risks associated with machine learning.

Taking the two conclusions together, the implication is that governance arrangements need to be reviewed. Two types of revision merit consideration. The first is that ways need to be found of balancing the benefits and risks of the use of machine learning tools. We observe that, in the main, discussions of benefits and risks are separate from one another, with enthusiasts in one bubble and those with concerns about privacy, confidentiality and other issues in another. They need to be brought together in some systematic and fruitful way. The second revision is that future governance arrangements will need to be less cybernetic, and more outward-looking. Citizens and other stakeholders do not have stakes in the arrangements, and cannot hold government actors to account. If trust in these organisations to manage personal data is at risk, then accounting to stakeholders will need to be taken seriously.

Acknowledgments

We are grateful to our project partners for their help and support, and to editors and reviewers for comments on an earlier version of this article. The project was funded by the Engineering and Physical

Sciences Research Council (EPSRC), EP/N013980/1.

References

Adnan, M., Nguyen, P., Ruddle, R. & Turkay, C. (2019) Visual analytics of event data using multiple mining methods. Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA), 2019.

Adnan, M. & Ruddle, R. (2018) A set-based visual analytics approach to analyze retail data. Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA), 2018.

AHSN Network (2020) Accelerating artificial intelligence in health and care: results from a state of the nation survey. https://bit.ly/30K8AHM – accessed 9th October 2020.

Badar, M., Haris, M. & Fatima, A. (2020) Application of deep learning for retinal analysis: a review. *Computer Science Review*, *35*, 100203.

Bardsley, M., Billings, J., Dixon, J. et al. (2011) Predicting who will use intensive social care: case finding tools based on linked health and social care data, *Age and Ageing*, *40*(2), 265-270.

Bates, D. (2015) Health information technology and care coordination: the next big opportunity for informatics? *Yearb Med Inform*, *10*(1), 11-4.

Boonstra, A., Versluis, A. & Vos, J. (2014) Implementing electronic health records in hospitals: a systematic literature review. *BMC Health Serv Res*, *14*, 370.

Bowker, G. & Star, S. (1999) Sorting Things Out. Cambridge MA: MIT Press.

Bycroft, C., Freeman, C., Petkova, D. et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*, 203-209.

Davis, S. (2017) The uncounted: politics of data and visibility in global health. *The International Journal of Human Rights*, *21*(8), 1144-63.

Dourish, P. & Cruz, E. (2018) Datafication and data fiction: narrating data and narrating with data. Big Data and Society, July-December 1-10.

Dourish, P. (2016) Algorithms and their others: algorithmic culture in context. Big Data and Society, July-December 1-11.

European Union (2016) General Data Protection Regulation.

Ezrachi, A., & Stucke, M. (2016) *Virtual competition: the promise and perils of the algorithm-driven economy*. Cambridge MA: Harvard University Press.

Fazlioglu, M. (2019) Beyond the "Nature" of Data: Obstacles to Protecting Sensitive Information in the European Union and the United States. *Fordham Urban Law Journal*, *46*(2), 271.

Fernow, J., de Miguel Beriain, I., Brey, P., & Stahl, B. (2019). Setting future ethical standards for ICT, Big Data, AI and robotics. *ORBIT Journal*, *2019*(1).

Future Advocacy (2018) Ethical, social and political challenges of artificial intelligence in health. London: Wellcome Trust.

Genomics England: https://www.genomicsengland.co.uk (accessed 9th October 2020).

Hanseth, O., & Ciborra, C. (Editors) (2007) Risk, complexity and ICT. Cheltenham: Edward Elgar.

Health and Social Care Information Centre (chair: Sir Nick Partridge) (2014) Data release review. Leeds: HSCIC.

Kaplan, B. (2016) How Should Health Data Be Used? *Camb Q Healthc Ethics*, *25*(2), 312-29.

Keen, J., Greenhalgh, J., Randell, R. et al. (2019) Networked information technologies and patient safety: a protocol for a realist synthesis. *Syst Rev*, *8*, 307.

Keen, J., Nicklin, E., Wickramasekera, N. et al. (2018) From embracing to managing risks. *BMJ Open*, *8*, e022921.

Kennedy, I. (1994) Between ourselves. *Br J Med Ethics*, *20*, 69-70.

Laurie, G., Stevens, L., Jones, K., & Dobbs, C. (2015) A Review of Evidence Relating to Harm Resulting from Uses of Health and Biomedical Data. London: Nuffield Council on Bioethics.

Malik, M. (2020) A Hierarchy of Limitations in Machine Learning. arXiv:2002.05193.

Manson, N., & O'Neill, O. (2007) Rethinking Informed Consent in Bioethics. Cambridge: Cambridge University Press.

Mittelstadt, B., & Floridi, L. (2016) The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics*, *22*, 303-41.

O'Hara, J., Aase, K., & Waring, J. (2019) Scaffolding our systems? Patients and families 'reaching in' as a source of healthcare resilience. *BMJ Qual Saf*, *28*(1), 3-6.

O'Neil, C. (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. London: Penguin.

O'Neill, O. (2002) Autonomy and trust in bioethics. Cambridge: Cambridge University Press.

Otero Varela, L., Wiebe, N., Niven, D. et al. (2019) Evaluation of interventions to improve electronic health record documentation within the inpatient setting: a protocol for a systematic review. *Syst Rev*, *8*, 54.

Palczewska, A. & Palczewski, J. (2019) Risk stratification for ASC services. Leeds: Leeds City Council.

Palczewska, A., Palczewski, J., Aivaliotis, G., & Kowalik, L. (2017) RobustSPAM for inference from noisy longitudinal data and

preservation of privacy. *Proceedings of 6th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

Pfoh, E., Abramson, E., Edwards, A. et al. (2014) The comparative value of 3 electronic sources of medication data. *American Journal of Pharmacy Benefits*, *6*, 217-24.

Pollock, N., & Williams, R. (2009) Software and organisations. London: Routledge.

Powles, J., & Hodson, H. (2017) Google DeepMind and healthcare in an age of algorthms. *Health and Technology*, *7*, 351-67.

Richardson, G. (1991) Feedback thought in social science and systems theory. Philadelphia, University of Pennsylvania Press.

Ruddle, R., & Hall, M. (2019) Using Miniature Visualizations of Descriptive Statistics to Investigate the Quality of Electronic Health Records. *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies – Volume 5: HEALTHINF*, 230-238.

Sacha, D., Sedlmair, M., Zhang, L. et al. (2019) What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing*, *268*, 164-75.

Selbst, A., Boyd, D., Friedler, S. et al (2019) Fairness and Abstraction in Sociotechnical Systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. *Atlanta: Association for Computing Machinery*, 59-68.

Shah, NH., Milstein, A. & Bagley, SC. (2019) Making Machine Learning Models Clinically Useful. *JAMA*, *322*(14), 1351-1352.

Tanner, A. (2017) Strengthening Protection of Patient Medical Data. New York: The Century Foundation.

Walker, J., Pan, E., Johnston, D. et al. (2005) The value of health care information exchange and interoperability. *Health Affairs*, *19*; 1.

Wiens, J., Saria, S., Sendak, M. et al. (2019) Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*, *25*, 1337-1340.