This is a repository copy of *'Data saves lives': ideational-material drivers of health data journeys in the UK*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/219489/

Version: Published Version

Original Research Article

'Data saves lives': Ideational-material drivers of health data journeys in the UK

Itzelle Medina-Perea[1] (iD), Jo Bates[1] (iD) and Andrew Cox[1]

Abstract

In this paper, we bring together the concepts of data valences and data journeys to examine how ideational and material factors work together to shape the movement of health data from the UK healthcare sector to universities for reuse in research. Specifically, we focus on the interaction of university-based researchers' constructs about data with the material conditions of health data circulation in the UK and how these dynamics drive greater circulation of health data through the data sharing infrastructure. Building on our empirical research, we identify four data valences or expectations about data present in the discourses of university-based researchers – vanguard, discovery, truthiness and actionability – and three material factors – investment in data, infrastructure and labour. We argue that the interaction of these factors has created a favourable environment for making data flow from the healthcare sector into the hands of university-based researchers. This work contributes to a better understanding of why health data reuse practices are expanding and being sustained, and it challenges previous health data reuse research that treats the drivers shaping data flows as self-evident or already determined.

Introduction

In recent years, all around the world, there has been growing interest in reusing health data originally collected to provide patient care for health research. There are an increasing number of initiatives that encourage data flows from the health sector to external bodies, and as a result, more data are flowing from the sector to other organisations to be reused. For example, in 2015, the US government launched the Precision Medicine Initiative; this plan included the allocation of funds to improve access to medical records for research (White House, 2015). In 2022, the European Commission established the European Health Data Space initiative, which proposes to facilitate the reuse of electronic health records for research purposes (European Commission, 2023). Initiatives like this have also been announced in countries such as Sweden (Government Offices of Sweden, 2020) and Finland (Ministry of Social Affairs and Health, n.d.). The same trend is found in the UK, with funding to support the reuse of health data growing and increasing numbers of research groups based at universities engaging in research reusing health data (UK Government, 2022a).

Researchers have explored health data reuse from different perspectives. It has often been taken for granted by researchers in the health data field that sharing data is a good thing; therefore, authors have tended to focus on barriers. For example, research has identified data quality and related issues that act as barriers to accessing and reusing health data (Edmondson and Reimer, 2020) and has suggested strategies to overcome such barriers (Sahoo et al., 2016). The role of public bodies in shaping barriers to the reuse of health data has also been explored (Aula, 2019), as have the perceptions of different groups of stakeholders, such as members of the public (Aitken, de St Jorre et al., 2016), patients (Tully et al., 2018) and healthcare professionals (Ford et al., 2020; Neves et al., 2019). This existing research on health data tends to pay little attention to the drivers underlying this growing circulation of data. Instead, the authors appear to bracket such drivers as already determined.

However, beyond this body of research, in fields such as critical data studies, researchers have paid more attention to the drivers of data circulation. Beer (2013), for example,

[1]Information School, The University of Sheffield, Sheffield, UK

Corresponding author:
Itzelle Medina-Perea, Information School, The University of Sheffield, Sheffield, UK.
Email: i.medinaperea@sheffield.ac.uk

identifies data circulation as co-constitutive of contemporary popular culture; Bates (2012) examines the neoliberal political economic agenda driving the Open Government Data agenda in the UK; and, in the education setting, Hartong and Foerschler (2019) examine the ways data circulation is incentivised through school performance targets that are used to justify funding. Some of this research has also addressed health data. For example, Van Dijck and Poell (2016) illuminate how commercial interests and technological change are driving health data to data brokers, and Vezyridis and Timmons (2019) identify the UK government's interest in the financial gains they perceive as an outcome of sharing NHS data. Similarly, in the Danish context, Hoeyer (2019) observed that the state's policy and fiscal challenges led to the growth of health data sharing initiatives.

This more critical body of work tends to draw attention to the political and economic drivers of data circulation in the context of broadly neoliberal governance regimes. This work has been vital in appreciating how ideational constructs about the economy and society contribute to the constitution of data circulation. However, despite widespread understanding that beliefs about data exist (Kitchin, 2014) and shape practice (Fiore-Gartland and Neff, 2015), researchers have tended not to examine in detail how ideational constructs about the actual data – in interaction with material conditions – also contribute to shaping its circulation. Understanding this dynamic is particularly important in the case of health data sharing where narratives about the power of data are widely promoted, although the insights gained from examining these dynamics are transferable to other settings.

Drawing on constructivist theories of political analysis (Hay, 2002), we can theorise that data-related ideational factors (e.g., beliefs, assumptions and expectations about data) in interaction with material conditions (e.g., funding, infrastructure and labour) are crucial in driving the circulation of health data for reuse for research purposes. Identifying and examining these interrelated drivers is important because it will help us to better understand how and why flows of health data are increasingly opening up towards universities, beyond political economic explanations, and why health data reuse practices are growing and being sustained. It will also challenge treating them as self-evident or already determined, as in the case of much-existing research on health data sharing. These insights are crucial in a context in which a trend of conducting research reusing health data is expanding. In this paper, we investigate health data flows in the UK context. Specifically, we focus on the interaction of the ideational milieu of university-based researchers with the material conditions of health data circulation in the UK and how these dynamics drive greater circulation of health data through the data sharing infrastructure.

To approach this, we draw on the concepts of data valences developed by Fiore-Gartland and Neff (2015) and data journeys developed by Bates et al. (2016). This is the first study that brings together these different strands of thinking to study the movement of data. Working with these concepts in combination allows us to make sense of how data-related ideational and material factors work together. While data journeys prompt us to consider how the sociocultural interacts with material conditions of production to shape and give life to data flows, data valences offer a way to further the data journeys conceptualisation of the sociocultural. Our data collection included 22 interviews with experts on the use and processing of health data and university-based researchers, as well as analysis of key documents produced by university-based research teams and key stakeholders in the UK healthcare landscape.

The paper is structured as follows. First, we introduce key insights from previous research on the reuse of health data and introduce the concepts of data journeys and data valences. We then explain our empirical methods, before going on to present the data-related ideational and material factors – and the interactions between them – that we identify as key to shaping the movement of health data in this context.

Barriers, attitudes and discourses

Within the existing health data sharing literature, three aspects of health data reuse have received significant attention: (a) barriers to reusing health data, (b) attitudes of different stakeholders and (c) government and scientific discourses.

Several studies that explore health data quality are consistent in pointing out that quality issues act as significant barriers to sharing and reusing health data (Edmondson and Reimer, 2020; Schlegel et al., 2017). Key findings of this body of work show that health datasets are often incomplete (Thiru et al., 2003), fragmented and contain inaccurate data (Kahn et al., 2012). This impacts data flows because if data has quality issues, additional work is needed to clean data before it can be reused which slows down workflows and potentially halts projects entirely. These issues have led some researchers to develop strategies or technical solutions to identify and address data quality barriers (Schlegel et al., 2017), such as the implementation of platforms or the application of statistics-based methods (Sahoo et al., 2016).

The second body of research, which explores the attitudes of different stakeholders towards the reuse of health data, has been primarily focused on understanding levels of support for data reuse outside the healthcare sector, conditions to support the reuse of data and awareness of and main concerns about data reuse practices. This work has been mainly motivated by an interest in understanding what uses of data are considered acceptable or appropriate and using these insights to inform policy (Aitken, Cunningham-Burley et al., 2016) and strategies to foster patient and public trust

in organisations conducting research with health data (Kalkman et al., 2019). Findings show that patients and members of the public tend to express positive attitudes about sharing health data with researchers (e.g. Aitken, Cunningham-Burley et al., 2016) and are supportive of this practice; however, this is not unconditional (Kalkman et al., 2019; Tully et al., 2018). Key conditions reported in the literature for acceptance from patients and the public in general are (a) that data are used for the public benefit (Skovgaard et al., 2019; Tully et al., 2018) and (b) that confidentiality is ensured (Aitken, de St Jorre et al., 2016; Kalkman et al., 2019).

Studies focused on the attitudes of healthcare professionals show that practitioners understand the benefits of health data research (Ford et al., 2020; Neves et al., 2019) and are willing to support it if adequate measures to handle data are in place and the interests of patients are protected (Ford et al., 2020). Nonetheless, healthcare professionals are likely to express concern and withdraw their support if they perceive that this could negatively impact patients or harm vulnerable groups (Neves et al., 2019).

Finally, previous research has also explored how health data has been framed in the discourses of authorities (Hoeyer, 2019) and healthcare and scientific editorials (Stevens et al., 2018). With an interest in better understanding the politics of health data initiatives, Hoeyer explored how promissory data interacts with notions about accountability in policies in the public health domain. He proposes that data are increasingly framed as a promise for future accountability. This means that in recent health data initiatives, authorities propose to wait for more data to be gathered to address health issues in the future rather than taking action in the present. This allows authorities to avoid or delay action to address health inequalities until more data are available. Hoeyer is critical of the position adopted by politicians regarding intensified data collection. On one hand, they are enthusiastic about collecting more data to achieve potential benefits of personalising medicine and prevention, but on the other hand, they fail to show the same level of enthusiasm towards using existing data when justifying the economic implications of their proposed investments.

Stevens and colleagues (2018), who explored the discourses about big data in healthcare in scientific editorials, identified five different discourses, which, according to them, emphasise certain aspects of big data while ignoring others. They observed that in the studied editorials, there is a strong presence of modernistic, instrumentalist and pragmatist discourses, which often reflect the idea that big data offers valid knowledge and that large-scale datasets and predictive analytics reflect the truth. Large amounts of data are often presented as objective, as an asset for an organisation that should not be questioned and as a resource that enables researchers to address previously unsolvable issues. While the authors also identified more critical discourses of big

data, namely, 'scientist' and 'critical-interpretive,' they observed that these have a weaker presence in editorials. Such discourses are characterised by questioning the objectivity and efficiency claims that dominate positive discourses about big data. These authors argue that the healthcare field would benefit from a more prominent critical-interpretive discourse (Stevens et al., 2018) that recognises that data do not speak for themselves, that social and political processes influence the creation of big data, and therefore, data are subjective, contain biases and have other limitations.

As the above review of existing research in the field of health data reuse suggests, little attention has so far been paid to the drivers shaping these health data flows. It is important to examine these drivers in the context of expanding the reuse of health data for research purposes. Our research addresses this gap by exploring how ideational and material factors work together to shape health data flows in the UK context.

Theoretical framework: Data journeys and data valences

While we used data journeys primarily as a methodological approach to guide research design (as elaborated on below), the concept is also a useful analytical lens to shed light on the socio-material dynamics shaping the circulation of data across different sites of practice, as well as the relational implications of these data flows. In this study, we were particularly interested in what data journeys can help to illuminate about the drivers of data flows. In the original data journeys paper (Bates et al., 2016), emphasis is placed on the materiality of data as a driver of circulation – specifically, the material properties of data such as its mutability, as well as the material conditions of production of data which act as drivers and frictions in data flows. These latter concerns about the material conditions of production are examined in depth in a further paper by the data journeys team examining the recovery and circulation of historic climate data (Bates et al., 2019). In our study, we continue with this focus. Rather than examining the material properties of data, we place emphasis on understanding how the material conditions of production drive the circulation of health data for research purposes. However, we are also interested in drawing out in more depth the 'sociocultural values' that Bates et al. (2016) argue combine with these material factors to shape data flows. In particular, we are interested in examining in detail how sociocultural beliefs and values (or ideational factors) about data interact with these material conditions of production to drive data circulation.

Since Bates et al. (2016) do not go into detail about how to unpack these ideational dynamics underlying data circulation, we can look to other frameworks for sociocultural analysis. Social imaginaries, for example, can be a powerful tool to understand collectively held narratives and

speculations expressed by people (Lupton, 2021), and with more focus on the technical, sociotechnical imaginaries can help to shed light on how powerful actors such as governments and public institutions imagine and shape technoscientific developments (Jasanoff and Kim, 2009). Elsewhere, methods such as critical discourse analysis can aid in understanding 'hidden relations and ideologies embedded in discourses…and examine the social and material consequences of discourse' (Johnson and McLean, 2020: 377).

'Data valences' is another such analytical lens and is the one that we adopt in our research. Fiore-Gartland & Neff (2015) define 'data valences' as the different expectations or value that people place on data in different contexts (such as their ability to offer explanations or solutions to problems). Based on their research in the health sector, they identified six data valences held by their research participants: self-evidence, actionability, connection, transparency, truthiness and discovery. These data valences act as mediators for data; that is to say, they play an important role in shaping 'the social and material performance of data,' including its movement across different contexts (2015: 1470). We chose data valences as an analytical lens for three reasons, beyond its relevance to examining the movement of data. Firstly, it brings into the picture not only people's discourses but also how they interact with data in specific contexts – their data practices. Secondly, data valences emphasise analysis of how 'strong' or 'prevalent' an expectation about data is in a given context, which is important for identifying key drivers of circulation. Third, it draws attention to the importance of understanding conflicts about data, such as tensions between institutions or stakeholders, and making sense of the dynamics that emerge when similar data valences appear in different contexts.

Working with the concepts of data valences and data journeys in combination allows us to make sense of how different factors work together to mediate the performance of data. Drawing on constructivist approaches to political analysis, we argue that neither the ideational nor the material factors lead the way; rather, they work together in interaction (Hay, 2002). It is, therefore, important to put equal weight on understanding each of the factors and how they interact with one another. In our case, this is to better understand the role they play in driving the flow of health data from the healthcare sector towards universities to be reused for research purposes; however, a similar form of analysis would also be valuable to apply in other contexts.

Methods

The data journey methodology (Bates et al., 2016) was used with some adaptations to explore the journeys of selected personal health data produced by the UK National Health Service (NHS) and reused for research in universities.

This study was conducted in two phases and explored how health data produced in the NHS travelled to different universities across the UK to be reused for academic research. Ethical approval was obtained from The University of Sheffield Research Ethics Committee.

In phase 1, conducted in 2018, our aim was to identify potential data journeys to explore and begin to make sense of how data produced in the healthcare sector flows to different sites of data reuse. We did this through desk research and four semi-structured interviews, of an average duration of 45 min, with experts familiar with the use and processing of health data within the health and care sector. Those interviewed were two senior members of staff from NHS Digital, one director of a health data-intensive research centre based at an academic institution and a senior data manager at a not-for-profit medical research organisation.

Based on our analysis of data in phase 1, we selected five data journeys to follow and conducted interviews with researchers working on each of them. The topics explored in these different projects were stroke prevention, antibiotic resistance, urinary tract infections in children, psychotic disorders and colorectal cancer.

In phase 2, conducted between 2018 and 2019, we focused our data collection on the final sites of practice in the identified data journeys, UK universities that reuse health data for research purposes. We collected key documents (e.g., annual reports, data sharing agreements, strategy documents and policy papers) produced by university-based research teams and key stakeholders in the UK healthcare landscape. We also conducted 18 semi-structured interviews of an average duration of 1 h, with researchers working on these projects. Examples of topics discussed in interviews included:

- Practices around access to health data, including data access request processes, interactions with data providers and data producers, etc.
- Motivations for working on projects reusing health data
- Challenges and drivers perceived concerning the reuse of health data
- Perceptions concerning the reuse of health data by different organisations and for different purposes (e.g., charities, pharmaceutical companies and university-based research teams)

All interviews were transcribed and given a code.

Thematic analysis (Braun and Clarke, 2006) was used to analyse the transcripts of interviews conducted in phase 2 and the documents collected through desk research in both phases. In order to keep the inductive coding and thematic analysis focused, the notion of data valences was used as a sensitising concept alongside data journeys. Drawing on Fairclough and Fairclough (2012), we also carefully examined 'emotive terms' and analysed metaphors or

frames because they play a key role in determining how people conceptualise their goals and circumstances and, therefore, how they act. We also analysed representations (such as statements, propositions or ideas) not as elements in isolation, but also considering the relationships and connections between representations and how they work together to form complete arguments.

Ideational and material drivers of health data journeys

The findings evidence how ideational constructs interact with material conditions of production to drive health data generated in the UK healthcare sector towards reuse in universities. This section starts by presenting four data valences driving data flows that were identified in the discourses of university-based researchers and key bodies in the UK healthcare data landscape. After this, we discuss three aspects of the material conditions. Finally, we examine how these data valences and material conditions interact to drive the flow of health data from the healthcare sector towards universities.

Data valences: Vanguard, discovery, truthiness and actionability

Below, we discuss the data valences, or expectations about data, that we identified as key drivers of health data circulation: vanguard, discovery, truthiness and actionability. The vanguard valence is a new valence proposed and defined by us, which emerged inductively from our thematic analysis. The three other valences identified – discovery, truthiness and actionability – were previously defined by Fiore-Gartland and Neff (2015). While other valences identified by Fiore-Gartland and Neff were found in our analysis, they were less relevant to the question of drivers of data circulation. For example, the connection valence which conveys the expectation that data can be an opportunity to connect people or engage in meaningful conversations was identified; however, it did not play a role in driving the circulation of data. Similarly, the self-evidence valence, which refers to the notion that data are 'pre-made' resources that require neither work nor interpretation, was identified, but not in a very strong form. Participants sometimes alluded to this valence; however, they also more commonly showed awareness of the complexities of working with data. They talked, for example, about the need to clean and organise data and run different rounds of analysis before being able to yield results. On the other hand, the transparency valence, which conveys the expectation of seamless flows of data across different contexts, certainly was present in a strong form in the discourses of university-based researchers. Researchers tended to hold an ideal vision that reflected the transparency

valence, and some of their activity was motivated by trying to achieve this vision. However, there were also more nuanced, and interesting, valences that were driving the circulation of data as discussed below.

Vanguard valence. We define the 'vanguard valence' as the expectation that conducting research with health data is the most innovative and cutting-edge way of exploring health issues. This valence differs from the valences defined by Fiore-Gartland and Neff in two key ways. Firstly, it captures the excitement and novelty associated with working with health data. Secondly, the vanguard valence highlights the expectation of potential outcomes, such as being positioned as a leader or pioneer in the field.

University-based researchers tended to evoke the vanguard valence when they talked about conducting research with large health datasets using data analytics techniques. They often highlighted that this type of research positions them as part of a group breaking with old ways of doing science.

Most participants across the different data journeys expressed the idea that they felt attracted to work in projects that involved the reuse of health data because this felt more novel than other types of research. For example, a member of the stroke project team commented:

> I wanted to do something more novel with the professors who are trying… to kind of work more at the cutting edge of innovation. I wanted to deal with new technologies that are changing with data we can collect, just getting a flavour of new, new developments. (S1–01)

Whereas interviewees talked about a broad range of techniques they have used to analyse data, the possibility of using predictive analytics seemed to generate more enthusiasm and excitement from some participants. Yet, the excitement for predictive analytics was not necessarily prompted by the results obtained using these techniques. Instead, the excitement was driven by the future potential that they saw in this type of analysis. For example, a participant working in the psychotic disorder project, aimed at predicting outcomes, talked with enthusiasm about the use of predictive analytics despite not knowing yet if successful results would be achieved in their project. They commented about how the attraction of the work lies in the possibility of achieving results in the future:

> … I do not know yet if it is possible to develop an accurate prediction model but let's say it is… that is the attraction, the potential that it could be used in real-world clinical settings. (S5–11)

Similarly, an early career researcher explained:

> I fell in love with tech, so my research involves mining medical records for early predictors of dementia… So,

I'm doing unsupervised analysis to kind of see where the patterns of the data during the whole 20 years before you actually are diagnosed with dementia. (S3–05)

Both accounts presented above are from researchers who were at an early stage of their work and did not know if they would be able to develop the prediction models intended. Rather, both of them thought it was worth being at the cutting edge of exploring their potential and felt excited about the prospect of success.

Researchers tended to speak with a sense of urgency about the need to continue conducting and incentivising health data-driven research, pointing out that university research should not be left behind other sectors that are already doing data-driven research. Although alluding to 'other sectors,' they mainly referred to big retail and technology companies such as Amazon and Google. These observations suggest that some researchers feel the need to keep up with other sectors they perceive to be at the vanguard.

The excitement of being at the vanguard, which is driving more people into conducting research with large amounts of health data, is not only present in the discourses of university-based researchers. The vanguard valence also underpins the agendas of key organisations that influence how data produced in the UK's healthcare sector are used. For example, the One Institute Strategy 2019/20, by Health Data Research UK (HDR UK), the UK's national institute for health data science, asserts that the population's health data will provide the 'UK's pioneers' (HDR UK, 2019c: 3) with an opportunity to 'revolutionise' (p. 7) healthcare. This strategy frames data as a resource that will give an opportunity to drive innovation, and a future scenario is presented –one in which the utilisation of health data science tools and technologies will result in radically improved healthcare.

Discovery and truthiness valences. The truthiness and discovery valences often appeared together in participants' accounts; therefore, it makes sense to discuss them in the same section. The truthiness valence illustrates the expectation that data offers a direct and objective representation of a measurable reality, while the discovery valence depicts how people expect data to be the source for discovering phenomena, issues, relationships or states that otherwise would remain unknown (Fiore-Gartland and Neff, 2015). Although appearing together, the discovery valence seemed to have a stronger presence than truthiness.

The presence of the truthiness valence was evident in that a key reason why participants believe that routinely collected patient data is a unique resource is because, for them, patient datasets generated in the UK healthcare sector are richer and more objective than datasets they otherwise might use:

These resources you can't really compare with anything else, so it is really rich and you got these long-term data. You could not really replicate that with a different type of dataset. (S3–05)

Identifying patterns in such datasets is what connects the truthiness to the discovery valence. As observed by Fiore-Gartland and Neff (2015), the discovery valence follows the logic that finding patterns in data is equal to knowing or understanding patterns in life whether at cellular, individual or population levels. The accounts of most university-based researchers interviewed reflected this logic. Across all data journeys, researchers tended to perceive that finding patterns in data was the same as knowing or understanding patterns in health at different population levels. For example, a participant expressed that:

I think there are lots of things you would not discover if you don't have access to these data. (S7–16)

A number of participants across all the data journeys commented that certain research questions could only be answered using routinely collected health data. For example, a researcher working on the antibiotic resistance project explained that this study allowed them to understand important aspects of antibiotic prescription, such as how antibiotics affected specific ethnic groups, and the rates of antibiotic prescribing in areas with different rates of deprivation. She pointed out that without these routinely collected health data, they would have found it very difficult to find answers to some of their questions:

using this kind of dataset you get the real-world view… it would be really hard to look at those questions without this dataset. (S6–13)

National organisations in the UK health research landscape have also expressed a key idea at the core of the discovery valence, which is that large health datasets have extraordinary power to enable discovery. For example, the HDR UK One Institute Strategy (HDR UK, 2019c) argues that through the mobilisation of molecular and clinical data, NHS routine data and data from wearable devices, the accelerated discovery of new diseases will be possible. However, in contrast to what we observed in relation to the discovery valence, the truthiness valence seems to have a weaker presence in the discourses of such organisations.

Actionability valence. A form of the actionability valence, which is defined as 'the expectation that data drive or do something within a social setting or that data can be leveraged for action' (Fiore-Gartland & Neff, 2015: 1474), was also identified in the discourses of university-based researchers.

When participants explained why they were interested in using health data, another reason they gave, in addition to those at the heart of vanguard, truthiness and discovery valences, was that 'data can save lives'. This phrase, which alludes to the actionability valence, was heavily used by university-based researchers and data practitioners and is also the name and slogan of a public engagement campaign originated at the University of Manchester's Health eResearch Centre in 2014 (Data Saves Lives, n.d.). This campaign has now been adopted by several research networks and stakeholder groups within the UK and other nations such as Australia and Denmark. Most university-based researchers and data practitioners interviewed for this research are part of research groups that have adopted the Data Saves Lives slogan, organised events, joined digital media campaigns and written case studies to gain public trust and support for health data reuse.

According to several participants' accounts, data saves lives because previously unsolved problems and unanswered questions can be addressed with data. Underlying these beliefs is the assumption that having knowledge or answers will lead to action. In relation to the research projects that they were involved in, university-based researchers tended to argue that health data leads to knowledge and that knowledge results in actions that positively impact the population's health.

The idea that data saves lives was expressed by participants across all the different data journeys explored, from early career to senior researchers. For example, one participant commented:

> I can tell you that yes, data save lives, and I think that in the academic and research environment we are, most of us are on board with that idea that data save lives. (S7–15)

The notion that 'data save lives' which, as stated above, evokes the actionability valence (Fiore-Gartland and Neff, 2015) was also identified in public discourses of key organisations and individuals in the UK healthcare sector landscape. For example, in a 2020 public appearance, Matt Hancock, the then Secretary of Health, alluded that 'data save lives':

> We have the biggest and most comprehensive health data system in the world in the NHS… data needs to be used to save lives. (Hancock, 2020)

This reveals that the expectation that data are actionable is prevalent not only within the community of university-based researchers but also in the discourses of other key players within the healthcare data landscape.

In this section, we discussed the different valences identified in the discourses of university-based researchers and how they are reflected in the discourses of key players in the sector, such as funders. Driven in part by the potential to conduct groundbreaking research, gain an accurate picture of population health, uncover patterns in health population through data analysis and save lives, researchers have been increasingly showing interest in taking up opportunities to reuse health data for research. We suggest that the assumptions conveyed through the vanguard, discovery, truthiness and actionability valences helped to create and reinforce a discourse that speaks in a positive tone about the reuse of health data for research. These valences in combination have helped to motivate and justify both the actions of researchers creating a growing demand for data and the key stakeholders with the power, visibility and resources to invest in setting the material conditions for fostering the movement of health data that are identified in the following section.

Material conditions of production: Investments in data, infrastructure and labour

In recent years, key stakeholders such as the UK National Health Service, the Department of Health and Social Care (DHSC) and the Medical Research Council (MRC) have invested in efforts to develop the material conditions for the production, use and circulation of health data. This includes three key developments: (a) investment in developing a secure data sharing infrastructure, (b) investment in increasing the quantity and quality of data available and (c) labour supply and provision of training in data science.

Investment in infrastructure for safe access and handling of health data. Data safe havens, also known as Trusted Research Environments (TREs), provide a suitable environment to access data generated in the healthcare sector securely within and beyond the NHS. The use of these data infrastructures is not new. However, although they have existed for a long time, until recently, the number of TREs available for researchers has been limited. In addition, as reported by our participants, even if researchers have access to data safe havens, technical issues such as occasional malfunctioning of these infrastructures when using them are not uncommon.

However, in recent years, the government has invested significantly in data safe havens to improve the capabilities of the existing ones, create more and support academic organisations in developing their own. These developments have facilitated university researchers' access to health data in a safe and secure way and helped to increase the demand for data flows towards universities.

The provision of more safe and secure data analytic facilities continues to be a priority for the government in the UK (UK Government, 2022b). Evidence of this is that the government announced in 2022 an investment of £200 million in national and regional TREs (Goldacre,

2022). The development of this secure data infrastructure is currently still in progress; however, through this investment, the DHSC expects to power 'life-saving research' (UK Government, 2022a: 7).

Investment in the quantity and quality of data. Another factor that has fostered the movement of health data towards universities is that investments aimed at increasing the amount and quality of health data available to university researchers have grown significantly in recent years. These investments and their outcomes are a further material factor driving up demand for more data to flow to researchers. This does not come entirely as a surprise as previous studies have highlighted that a factor contributing to fostering data circulation is the greater availability of data (De Roo et al. 2016).

For example, in 2019, HDR UK announced the launch of the Digital Innovation Hub (DIH) Programme financed by the UK Research and Innovation's Industrial Strategy Challenge Fund (a £37.5 million investment) (HDR UK, 2019a). When this initiative was launched, nine health data research hubs were created with the aim of providing a rich toolkit of healthcare datasets, infrastructure and capabilities that enable users to 'identify, access, understand and use data' (HDR UK, n.d.-b: 3). These hubs have contributed to increase the amount of health data available because one of their main efforts consisted of curating, improving and linking datasets so they are ready to use. As a result of this, more than 200 datasets have been made available for research, and there are plans to increase this number over the coming years (HDR UK, 2022).

More recently, the government strategy on the use of NHS data, *Data saves lives: reshaping health and social care with data* (UK Government, 2022a), identified the HDR UK hubs as one of their most successful programmes and indicated the government's interest in continuing to foster efforts to increase the scale and quality of data sets available to researchers (UK Government, 2022a).

According to university-based researchers, the large amount of data available is, in itself, a reason to use it, although they did not explicitly link this to investments from the government. For example, one participant commented:

> the data is there…we have never had access to electronic health records in this way before, so the fact that that resource is there, I think is in itself a driving factor for using it. (S5–11)

The availability of data as a result of these investments in data quantity and quality is therefore a key material factor driving data through the infrastructure and into the hands of researchers.

Recognising that lack of quality metadata is a key barrier to finding and using health data, some groups of researchers have undertaken metadata improvement projects to develop detailed descriptions of data that are available from different sources. This is not only to be able to use the metadata in

their own projects but also to facilitate the access and use of datasets for other researchers.

Employment opportunities and the investment in training to develop data science skills. The capacity of the UK's labour force to work with health data is also a key factor of the material conditions of production which has helped to drive the movement of health data towards universities. The MRC and other organisations launched HDR UK in 2018, to 'unite the UK's health data to allow discoveries that improve the lives of people' (HDR UK, n.d.-c: 1). In its first 5 years, HDR UK aimed to train more than 10,000 health data scientists (HDR UK, 2019b) and support them in becoming leaders within health data research and with these efforts, 'lead the health data science revolution' (HDR UK, 2019b: 10). As part of its strategy to achieve its goals, HDR UK launched the fellowship programme New Leaders in Health Data Research. Fellows of this programme benefit from 3 years of funding and what the institute has labelled as 'cutting-edge' data science training, mentoring support and leadership opportunities (HDR UK, n.d.-a). According to the institute, a number of fellows have established their own research groups and obtained funding from various sources to train 'future generations of health data scientists' (HDR UK, n.d.-a: 4).

The material benefits (in terms of financial and career rewards) of engaging in health data research are also something university-based researchers are acutely aware of. When talking about their work, they commented on how using health data for research generates career opportunities in UK academia. For example, one interviewee commented:

> [Health data] is just providing a job for researchers. (S4–07)

University-based researchers across all the different projects explored perceived that conducting research with health data gives them the opportunity to produce high-profile academic papers, which could help them advance their academic careers. This view was shared among participants of this study, from early career researchers to senior academics. A researcher commented:

> From a cynical perspective, academics are interested in advancing their careers, so they want those high-profile papers. (S4–07)

Researchers showed awareness of these material benefits, and this has, in part, motivated them to continue working with health data. The fact that the government has prioritised investing in initiatives to allow researchers to develop or improve data science skills and supporting health data research has contributed to foster the movement of health data from the healthcare sector towards the hands of university-based researchers, and researchers have

welcomed these initiatives. Researchers have been motivated to work with patient data in part because they embrace expectations at the heart of the vanguard, discovery, truthiness and actionability valences, as earlier discussed. But, also, they have been motivated to continue working with data because they recognise the material benefits that these initiatives bring, particularly in a context where one of the key factors used to evaluate their performance is their publishing record, and their career progress is dependent on this.

Interactions between data valences and material conditions of production

The above findings point to some of the ways that data valences and material conditions of production drive the movement of health data from the healthcare sector to universities to be reused for research purposes. However, drawing on constructivist theories of political analysis (Hay, 2002), we understand that these two types of factors do not work in isolation but work together to drive up demand for data among university-based researchers and shape the movement of health data. In other words, neither the ideational nor the material factors lead the way; rather they work together in interaction. Below, we analyse how the ideational and material factors we observed interact to drive the movement of health data.

In recent years, the government has invested in the improvement of data sharing infrastructures that support the reuse of health data – more specifically, the improvement of existing data safe havens and the creation of new ones. Key arguments that have been used to justify these investments allude to the discovery and vanguard valences. The former is projected through the notion that health data has an incredible potential to enable discoveries to 'save lives' (UK Government, 2022a), while the latter through the notion that by facilitating the access to research-ready data to researchers, these data sharing infrastructures help to drive the most 'cutting-edge research' (HDR UK, 2019b). These changes in the material conditions of data-driven health research have deepened researchers' beliefs and expectations as conveyed in the vanguard, discovery, truthiness and actionability valences, which also have a strong presence in the discourses of key players in the UK healthcare data landscape. Prompted, in part by these assumptions, university-based researchers have increasingly engaged in data-driven research and therefore made use of these data sharing infrastructures, which in turn has led to more investment in data sharing infrastructures. These dynamics, therefore, have helped to foster the movement of health data towards universities.

At the same time, the quality of datasets has also been improving, again as a result of government investments in large-scale initiatives supported by multiple bodies aimed at elevating the quality of health datasets and making them available for research. Examples of such initiatives include the Digital Innovation Hub Programme (HDR UK, 2019a) and the Data For Research Development programme (UK Government, 2022a). These changes in the quality and amount of data available for data-driven health research have been both motivated by and deepened researchers' beliefs and expectations depicted in the discovery and actionability data valences.

The growing availability of datasets has been welcomed by researchers, given that they hold expectations of data being actionable, able to offer a direct representation of a measurable reality, and the source through which they can discover phenomena and issues that otherwise would remain unknown, and they have been requesting access to the datasets available and making use of them to conduct data-driven research. Moreover, these expectations projected in the actionability, discovery and truthiness valences have also motivated some researchers to engage in data quality improvement projects to 'help to illuminate to other researchers what is available' (S1–01). Although at a smaller scale compared with governmental efforts, these projects have contributed to increasing the quality and quantity of data available for research. These dynamics feed one another; the availability of more quality data drives growth in the demand (powered by the expectations and beliefs depicted in the discovery, truthiness and actionability valences), which leads to an increased investment in data quality efforts. This results in a constantly expanding pool of data, which fuels even more demand for data. The cycle continues as both the demand and availability of quality health data keep growing.

Finally, in recent years, we have seen a consistent and growing stream of funding aimed at providing training to develop data science skills, support researchers in becoming leaders of new generations of data scientists and finance data-driven research projects which reuse health data thus providing employment opportunities for researchers working in this area. This interacts with beliefs and expectations at the core of the vanguard valence. Thus, the existence of this material support, combined with the expectation projected in the vanguard data valence, that using health data to conduct data-driven research is the most innovative and cutting-edge way of exploring health issues, has led researchers to engage in training to gain relevant abilities to build a career in this field. It has also allowed initiatives that seek to expand the number of people with key skills for data-driven research to thrive. Not only has the number of people equipped with skills to conduct data-driven research grown, but more people are taking up leading positions as ambassadors and encouragers of others to join them in the field of data-driven research. Up to March 2021, HRD UK had trained 6074 health data scientists and developed 46 fellows, whom they referred to as a 'new generation of 46 leaders in the [health data research] field'. These leaders contribute by

leading training programmes for health data scientists, promoting the importance of health research, and establishing collaborations for health data research at national and international level (HDR UK, n.d.-a). As can be seen, the big government investment in funding training to develop data science skills and the fact that there is a labour shortage at present in this field feed the belief that reusing health data to conduct data-driven research is the most innovative way of exploring health issues and working in this area gives the opportunity for researchers to be at the vanguard. This interaction has helped to foster the movement of health data because with more people trained in this area, growing employment opportunities in this field and more people acting as leaders and encouragers for others to join this field, more people are applying for access to the growing pots of data available and more people are using the data sharing infrastructure.

Smoother flows of data from the healthcare sector to universities are therefore the result of interactions between ideational and material factors. That is to say, in line with constructivist political analysis (Hay, 2002), neither the ideational nor the material factors lead the way; rather they work together in interaction. The dynamics described above have powered a cycle in which we observe better infrastructure, availability of larger amounts of data with higher quality, a growing number of people interested in building a career in data-driven research and therefore a good response to calls to use more data and improve skills. Thus, more investment and resources are directed to make data-driven research thrive, and again more people working in this area, and a constant reaffirmation that this is the best path to follow.

Discussion

This study contributes by enhancing the understanding of how material conditions of production and ideational constructs about data work together to shape the movement of health data in the UK healthcare sector. Previous research about health data reuse has extensively explored factors that play a role in creating barriers to the movement of health data as well as perceptions of different stakeholders regarding the reuse of these types of data for research purposes. Research in the critical data studies field has tended to consider, in broad terms, the political and economic drivers of data circulation. However, we knew little about how ideational constructs about data and material conditions of production interacted to contribute to driving the circulation of data, in our case health data for reuse for research purposes. We have demonstrated that the expectations that people have for data play a vital role in mediating the movement of data, but only in interaction with the material conditions for data circulation.

A shared characteristic of the valences we identify is that they all frame conducting research with large datasets in an unproblematically positive way; contrasting negative valences about using large amounts of health data for research were not identified. This, to a certain extent, is similar to what Stevens et al. (2018) identified in their exploration of the discourses of editorials in healthcare domain publications. Doing research using large volumes of patient data can lead to positive results, and there is some evidence that in certain areas this mass of data is helpful to the production of original and important health research (Wellcome Trust, 2015). However, embracing the grand promises of data expressed in the data valences is not unproblematic. University-based researchers and data practitioners often connected doing research with large patient datasets to ideas such as innovation and framed this type of research as a possibility to radically transform healthcare. Adopting this position has the risk that they can become so mesmerised with the power and promises data offers, that appreciating its limitations can become challenging (Mayer-Schönberger and Cukier, 2013). This could be seen as the risk of placing one's research in the 'vanguard'. Another risk is that embracing the promises of big data can lead to apophenia, which is seeing patterns in places where they do not exist, only because large volumes of data can provide associations that point in different directions (Leinweber, 2007). Data without theory can probably not lead to knowledge construction, challenging an assumption in the discoverability valence. Equally, finding patterns in data is not equivalent to having a better understanding of health issues at population scales, especially for minoritised populations who are consistently under-represented in health data. Finally, knowing the answers to research questions on its own does not lead to addressing health issues on the ground, so the logic of actionability also needs to be queried. This suggests that a more nuanced approach towards the value of working with large amounts of data is needed.

We have observed that the interaction of data valences with existing and emergent material conditions has created a favourable environment for making data flow from the healthcare sector to the hands of university-based researchers. University-based researchers have tended to embrace the big promises of big data. Their discourses and those pronounced by bodies such as the HDR UK seemed to be aligned. Key bodies in the healthcare sector often talk about health data as powerful assets, arguing that as more data are collected, shared and reused for research, we increase the possibilities of reaching a future where critical health issues can be addressed, and the health of the population can be improved. Some of the ideas at the core of the HDR UK strategy resemble the arguments at the core of data initiatives that according to Hoeyer (2019), frame data as a promise for future accountability and where intensified data collection is presented as leading the way towards public authorities being able to solve problems that cannot be addressed at the present time.

In the first instance, we argue that these organisations' discourses have helped reinforce and foster the ideas of researchers about the benefits of conducting research using health data. But beyond this, the big promises of big data have also underpinned the agendas of these organisations and have been presented as justifications to take various actions in an attempt to foster the reuse of health data. At the same time, we have observed an increasing demand for data stemming from research groups based at UK universities, which seems to be not only validated but also backed up by institutions with financial and material resources, visibility and decision-making power. For example, as previous research pointed out, data quality issues have been a barrier to the movement of health data (Edmondson and Reimer, 2020). We observed how the expectations and beliefs that people hold about data, conveyed through the discovery and actionability valence in interaction with investment in the quantity and quality of data, have worked together to ease such a long-standing barrier and drive the movement of health data from the healthcare sector towards universities.

The interaction of expectations of university-based researchers with material conditions of production has fostered the movement of health data towards universities for research purposes. Of course, this has enabled the development of research projects which have had a positive impact and have contributed to improving people's lives, such as the Building Rapid Interventions to Reduce Antibiotic Resistance project, conducted by a team of researchers at the University of Manchester, which contributed to the optimisation of antibiotic usage in primary care (Connected Health Cities, n.d.). At the same time, this has helped promissory health data strategies (Hoeyer, 2019) in the UK to continue to exist and grow.

Conclusion

In a context where narratives about the power of data are widely promoted, such as in the case of health data sharing, it is crucial to understand how different factors contribute to shaping the circulation of data. Previous research in the critical data studies field sheds light on how ideational constructs about the economy and society contribute to the constitution of data circulation. Furthermore, it is widely understood that beliefs about data are significant (Kitchin, 2014) and shape practice (Fiore-Gartland and Neff, 2015); however, we knew little about how ideational constructs about the actual data – in interaction with material conditions – also contribute to shaping their circulation.

The key contribution of this paper is in bringing together the concepts of data valences (Fiore-Gartland and Neff, 2015) and data journeys (Bates et al., 2016) to examine the movement of data. Data journeys prompted us to consider how the sociocultural interacts with material conditions to shape data flows, while data valences offered us a

way to further the Data Journeys conceptualisation of the sociocultural. Working with these concepts in combination allows us to understand how ideational constructs about data and the material conditions of production interact to drive the flow of health data from the UK healthcare sector to universities for reuse in research, and a similar approach is likely to shed light on drivers of data circulation in other contexts. Drawing on constructivist approaches to political analysis (Hay, 2002), we have argued that data-related ideational factors (e.g., beliefs, assumptions and expectations about data) in interaction with material conditions (e.g., funding, infrastructure and labour) work together to shape the circulation of health data for reuse for research purposes. However, neither of them necessarily determines the outcome. Therefore, it is essential to give equal attention to understanding each of the factors and how they interact with one another. This work also contributes to a better understanding of why health data reuse practices are expanding, and it challenges the notion of treating the drivers shaping data flows as self-evident or already determined, as in the case of much health data reuse research. Findings demonstrated that the expectations that people have for data play a vital role in mediating the movement of data, but only in interaction with the material conditions for data circulation. We have observed that the interaction of vanguard, discovery, truthiness and actionability data valences with existing and emergent material conditions, namely, investment in developing a secure data sharing infrastructure, investment in increasing quantity and quality of data available and labour supply and provision of training in data science, have created a favourable environment to making data flow from the healthcare sector to the hands of university-based researchers. These dynamics have fuelled a cycle where we observe enhanced infrastructure, increased availability of higher-quality data, more people interested in pursuing a career in data-driven research, and a positive response to calls for greater use of data and skills improvement. This has led to further investment in data-driven research, attracting more people to work in this area and a reaffirmation that this is the most promising path. Another contribution of our work is the definition of an additional data valence distinct from those outlined by Fiore-Gartland and Neff (2015). Emerging inductively from our thematic analysis, we defined the 'vanguard valence' as the expectation that conducting research with health data is the most innovative and cutting-edge way of exploring health issues. This new valence captures the excitement and novelty associated with working with health data and highlights the expectation of potential outcomes, such as being positioned as a leader of pioneer in the field.

University-based researchers and other significant players in the UK health data landscape, such as funders, may be particularly interested in these findings. They can aid in their critical reflection on their own practices,

discourses and decision-making, the factors driving data movement and who gains and loses from the growing flow of health data from the healthcare sector to universities for reuse. While this study focused on the circulation of health data for reuse purposes in the UK context, future research could explore what and how ideational and material factors interact to shape the movement of health data in other countries, offering interesting and novel comparative insights. Applying a similar approach to other types of data and contexts could also produce new findings about drivers of data circulation.

Acknowledgements

Declaration of conflicting interests

Funding

ORCID iDs

Itzelle Medina-Perea 🔟 https://orcid.org/0000-0003-1702-1484
Jo Bates 🔟 https://orcid.org/0000-0001-7266-8470

References

Aitken M, Cunningham-Burley S and Pagliari C (2016) Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme. *Science and Public Policy* 43(5): 713–723.

Aitken M, de St Jorre J, Pagliari C, et al. (2016) Public responses to the sharing and linkage of health data for research purposes: A systematic review and thematic synthesis of qualitative studies. *BMC medical Ethics* 17(1): 73.

Aula V (2019) Institutions, infrastructures, and data friction – reforming secondary use of health data in Finland. *Big Data & Society* 6(2): 205395171987598.

Bates J (2012) "This is what modern deregulation looks like": co-optation and contestation in the shaping of the UK's Open Government Data Initiative. *The Journal of Community Informatics* 8(2).

Bates J, Goodale P, Lin Y, et al. (2019) Assembling an infrastructure for historic climate data recovery: Data friction in practice. *Journal of Documentation* 75(4): 791–806.

Bates J, Lin Y-W and Goodale P (2016) Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society* 3(2): 1–12.

Beer D (2013) *Popular Culture and New Media – The Politics of Circulation*. Palgrave Macmillan.

Braun V and Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2): 77–101.

Connected Health Cities (n.d.) BRIT – Using data to tackle antibiotic resistance – Connected Health Cities. Available at: https://www.connectedhealthcities.org/research-projects/using-data-tackle-antibiotic-resistance/ (accessed 25 February 2021).

Data Saves Lives (n.d.) #DataSavesLives. Available at: https://datasaveslives.eu/datasaveslives (accessed 15 November 2022).

De Roo B, De Maeyer P and Bourgeois J (2016) Information flows as bases for archeology-specific geodata infrastructures: An exploratory study in Flanders. *Journal of the Association for Information Science and Technology; Hoboken* 67(8): 1928–1942.

Edmondson ME and Reimer AP (2020) Challenges frequently encountered in the secondary use of electronic medical record data for research. *Computers, Informatics, Nursing: CIN* 38(7): 338–348.

European Commission (2023) European Health Data Space. Available at: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (accessed 16 February 2023).

Fairclough I and Fairclough N (2012) *Political Discourse Analysis: A Method for Advanced Students*. London: Routledge.

Fiore-Gartland B and Neff G (2015) Communication, mediation, and the expectations of data: Data valences across health and wellness communities. *International Journal of Communication* 9: 1466–1484.

Ford E, Kazempour Y, Cooper MJF, et al. (2020) Media content analysis of general practitioners' reactions to care.data expressed in the media: What lessons can be learned for future NHS data-sharing initiatives? *BMJ Open* 10(9): 1–9.

Goldacre B (2022) Better, broader, safer: using health data for research and analysis. Available at: https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis/better-broader-safer-using-health-data-for-research-and-analysis.

Government Offices of Sweden (2020) Sweden's national life sciences strategy. Available at: https://www.government.se/information-material/2020/11/swedens-national-life-sciences-strategy/ (accessed 16 February 2023).

Hancock M (2020) #LTW 2020. Founders Forum HealthTech Summit: Keynote from Secretary of State for Health. Video. London. Available at: https://www.youtube.com/watch?v=2gXnC57RK2Q (accessed 12 December 2020).

Hartong S and Foerschler A (2019) Opening the black box of data-based school monitoring: Data infrastructures, flows and practices in state education agencies. *Big Data & Society* 6(1): 1–12.

Hay C (2002) *Political Analysis: A Critical Introduction*. London, United Kingdom: Bloomsbury Publishing Plc. Available at:

http://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=4762648 (accessed 14 May 2023).

HDR UK (2019a) HDR UK Digital Innovation Hub Programme. Available at: https://www.hdruk.ac.uk/wp-content/uploads/2019/07/Digital-Innovation-Hub-Programme-Prospectus.pdf.

HDR UK (2019b) News Announcement: Health Data Research UK releases One Institute Strategy. Available at: https://www.hdruk.ac.uk/news/health-data-research-uk-releases-one-institute-strategy/ (accessed 21 December 2021).

HDR UK (2019c) One Institute Strategy 2019/20. HDR UK. Available at: https://www.hdruk.ac.uk/wp-content/uploads/2019/11/191010-HDR-UK-One-Institute-Strategy-compressed-for-website.pdf.

HDR UK (2022) HDRUK Health Data Research UK: Annual Review 2021/2022. Available at : https://www.hdruk.ac.uk/wp-content/uploads/2022/10/Annual-Review-Final-2021-22.pdf.

HDR UK (n.d.-a) HDR UK Fellows: the UK's new leaders in health data research. Available at: https://www.hdruk.ac.uk/careers-in-health-data-science/science-community-and-partners/fellowships/ (accessed 1 February 2023a).

HDR UK (n.d.-b) Health Data Research Hubs. In: HDR UK. Available at: https://www.hdruk.ac.uk/helping-with-health-data/health-data-research-hubs/ (accessed 1 February 2023b).

HDR UK (n.d.-c) What we do. Available at: https://www.hdruk.ac.uk/about-us/what-we-do/ (accessed 3 December 2020c).

Hoeyer K (2019) Data as promise: Reconfiguring Danish public health through personalized medicine. *Social Studies of Science* 49(4): 531–555.

Jasanoff S and Kim S-H (2009) Containing the atom: Sociotechnical imaginaries and nuclear power in the United States and South Korea. *Minerva* 47(2): 119–146.

Johnson MNP and McLean E (2020) Discourse analysis. In: Kobayashi A (eds) *International Encyclopedia of Human Geography*, Second Edition. Oxford: Elsevier, 377–383.

Kahn MG, Raebel MA, Glanz JM, et al. (2012) A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical Care* 50(7): S21–S29.

Kalkman S, van Delden J, Banerjee A, et al. (2019) Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *Journal of Medical Ethics* 48(1): 3–13.

Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Sage Publications Ltd.

Leinweber D (2007) Stupid data miner tricks: Overfitting the S&P 500. *The Journal of Investing* 16: 15–22.

Lupton D (2021) 'Not the real me': Social imaginaries of personal data profiling. *Cultural Sociology* 15(1): 3–21.

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Ministry of Social Affairs and Health (n.d.) Secondary use of health and social data. Available at: https://stm.fi/en/secondary-use-of-health-and-social-data (accessed 16 February 2023).

Neves AL, Poovendran D, Freise L, et al. (2019) Health care Professionals' perspectives on the secondary use of health records to improve quality and safety of care in England: Qualitative study. *Journal of Medical Internet Research* 21(9): e14135.

Sahoo SS, Zhang G-Q, Bamps Y, et al. (2016) Managing information well: Toward an ontology-driven informatics platform for data sharing and secondary use in epilepsy self-management research centers. *Health Informatics Journal* 22(3): 548–561.

Schlegel DR and Ficheur G (2017) Secondary use of patient data: Review of the literature published in 2016. *Yearbook of Medical Informatics* 26(1): 68–71.

Skovgaard LL, Wadmann S and Hoeyer K (2019) A review of attitudes towards the reuse of health data among people in the European union: The primacy of purpose and the common good. *Health Policy (Amsterdam, Netherlands)* 123(6): 564–571.

Stevens M, Wehrens R and de Bont A (2018) Conceptualizations of big data and their epistemological claims in healthcare: A discourse analysis. *Big Data & Society* 5(2): 2053951718816727.

Thiru K, Hassey A and Sullivan F (2003) Systematic review of scope and quality of electronic patient record data in primary care. *BMJ* 326(7398): 1070.

Tully MP, Bozentko K, Clement S, et al. (2018) Investigating the extent to which patients should control access to patient records for research: A deliberative process using Citizens' juries. *Journal of Medical Internet Research* 20(3): e112.

UK Government (2022a) Data saves lives: reshaping health and social care with data. Available at: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data (accessed 15 November 2022).

UK Government (2022b) Health Secretary sets out ambitious tech agenda. Available at: https://www.gov.uk/government/news/health-secretary-sets-out-ambitious-tech-agenda (accessed 1 February 2023).

Van Dijck J and Poell T (2016) Understanding the promises and premises of online health platforms. *Big Data & Society* 3: 1.

Vezyridis P and Timmons S (2019) Resisting big data exploitations in public healthcare: Free riding or distributive justice? *Sociology of Health & Illness* 41(8): 1585–1599.

Wellcome Trust (2015) Enabling Data Linkage to Maximise the Value of Public Health Research Data: final report to The Wellcome Trust. Available at: https://wellcome.org/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf (accessed 14 May 2023).

White House (2015) Precision Medicine Initiative. Available at: https://obamawhitehouse.archives.gov/node/333101 (accessed 16 February 2023).