The University Of Sheffield.

Socio-material factors shaping patient data journeys in the United Kingdom

Itzelle Aurora Medina Perea

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Social Sciences
Information School

February 2021

Acknowledgements

Many thanks to the Mexican National Council for Science and Technology (CONACYT) for providing me with the funding to complete this PhD through the scholarship under register number CVU 692534.

Embarking in the PhD journey was a very challenging experience. During these four years I received encouragement and support from many people. In particular, I am extremely grateful to my supervisors Jo Bates and Andrew Cox who were very generous with their time and always provided insightful suggestions. They inspired and motivated me, but also asked me the most challenging questions and pushed me to go beyond my limits. Whenever I felt lost, they helped me to find direction.

I was very lucky to be part of the Information School, were I always felt welcomed and valued. Special thanks to Sukaina Ehdeed for her practical suggestions, encouraging words, and friendship. Particularly helpful to me during this journey were Harriet Godfrey, Peter Stordy, Andy Stones, Briony Birdy and Jorge Martins, who were always warm and friendly. Thanks should also go to Andrea Jimenez and Susan Oman for their emotional support and solidarity.

I would also like to extend my gratitude to my friends Esther Coello, Mayra Sanchez, and Sandra Zamudio for walking by my side through the ups and downs of this journey.

The completion of my thesis would not have been possible without the support of my mum, Aurora Perea, and my brother, Ernesto Medina. Their encouragement and profound belief in my abilities gave me strength to keep going.

Finally, I wish to thank my beloved León Felipe. He patiently listened to me whenever I needed to speak out loud to make sense of my ideas, encouraged me to take healthy breaks, wiped my tears in my darkest days, and celebrated my achievements. These years were a lot more colourful thanks to him.

Abstract

In a context where recent controversial data sharing initiatives between the NHS and external parties have generated dissatisfaction, confusion and uneasiness among the public, it is relevant to pay attention to social, cultural, political and ethical issues related to how power dynamics play out in the production, use and movement of patient data. Adopting a Critical Data Studies perspective, this in-depth qualitative study contributes to understanding how sociocultural values interact with material conditions to shape "data journeys" in the UK health sector, focusing on the reuse of patient data for research. By applying the Data Journeys approach (Bates et al., 2016) and integrating the notion of Data Valences (Fiore-Gartland & Neff, 2015) this work helps to understand university-based researchers' culture, how it relates to other data cultures, and how that impacts the flow of patient data generated in the UK healthcare sector.

In order to address the research aim, a critical thematic analysis of interviews with university-based researchers and other key stakeholders in the field, and key documents (e.g. data sharing policies, legislation, annual reports) was conducted. Five data journeys were followed as data travelled to be used in projects exploring stroke prevention, antibiotic resistance, urinary tract infections, psychotic disorders, and rectal cancer.

Findings of the research indicate that university-based researchers have tended to embrace the big promises of big data and their views seem to be in harmony with those of funders, policymakers and some data providers. Researchers tend to perceive themselves as a virtuous group that conducts research to benefit the UK population, and who therefore should have access to patient data without frictions. Some groups of patients and the public have endorsed this self-identity. These factors have generated a growing and validated demand for data for research purposes, which has helped to drive the flow of patient data towards universities. The endorsement of policymakers, funders and data providers has become material in the provision of resources such as funding and infrastructure while the public support has prevented the emergence of opposition towards their work.

The decisions of information governance staff act as frictions, as they have sometimes refused to share data, not accepting the virtuous self-identity of researchers as a justification to allow smooth data flows. Despite technology advances and investments, there are still infrastructural barriers and data management practices that slow down or block the movement of patient data to universities. While researchers recognise these factors can act as barriers, their belief in the

promises of big data motivates them to try to improve the quality of datasets and solve infrastructural issues that act as frictions for the movement of data towards them.

This study helps to understand the sociocultural values and norms within university-based research teams and how they interact with material elements to generate the socio-material conditions that create drivers and frictions that work together to shape the journeys of patient data reused for research purposes.

This study invites researchers to recognise that datasets can never offer accurate representations of the health of the country and to recognise themselves as political actors and adopt a reflective stand towards their own practices and discourses.

Table of Contents

List of figures

List of tables

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously been presented for an award at this, or any other, university.

Conference papers arising from this research:

Medina Perea, I. A., Bates, J. & Cox, A. (2019) Using data journeys to inform research design: Socio-cultural dynamics of patient data flows in the UK healthcare sector. iConference 2019 Proceedings

Medina Perea, I. A., Cox, A., & Bates, J. (2020). Exploring the life of patient data in the UK healthcare sector. AoIR Selected Papers of Internet Research, 2020.

Chapter 1. Introduction

1.1 Background and significance

In the United Kingdom (UK), the National Health Service (NHS) is responsible for delivering public healthcare, setting the priorities and direction for the whole National Healthcare System, and promoting and informing the national debate to improve health and care (NHS Choices, 2016). To date, the NHS is largest publicly funded health service in the world (White, 2010), and most of the residents of the UK are NHS patients (Department of Health & Social Care, n.d.). Amongst the different types of data that are produced in the NHS, we can find patient data.

In "the datafication era" the nature of data and the ways in which data are generated, processed, analysed, stored, and shared is being drastically transformed (Kitchin, 2014d). In the healthcare sector, personal patient data, which are defined as data and information extracted from patient records, are collected, processed and used in different ways (NHS Choices, 2016). These data can be utilised for different purposes (Understanding Patient Data, 2017), and are moved across different sites of practice in response to different information needs. Patient data may include details about medical conditions; notes recorded by doctors, nurses, or other healthcare professionals; and personal details such as NHS number and date of birth (Connected Health Cities, 2016; Understanding Patient Data, 2017). This term may be used to refer to a whole record or only a part of it. Patient records include data relating to the individual's health (mental and physical) and the individual's social care needs and services (Connected Health Cities, 2016). Specifically, medical records of the NHS can include, for example:

- Medical test results (i.e. allergy tests, blood tests and other screenings)
- General details such as name, address and age
- Consultation notes taken by a doctor during an appointment
- Results of treatment
- Medicines that a patient takes
- Details about long-term conditions of a patient such as asthma or diabetes
- Clinically relevant lifestyle details, such as alcohol consumption, smoking, or weight.

In the UK healthcare sector, the primary use of personal patient data is to provide care to patients (NHS Digital, 2018d). To be more specific, these data are used with the objective of

protecting, promoting, maintaining or meeting the health needs of a patient (Health Information and Quality Authority, 2012). However, this type of data is increasingly also being used for other purposes (Health Information and Quality Authority, 2012; NHS Digital, 2018d). Examples of such secondary uses of patient data include healthcare planning, creation of policy, and research in academic institutions (NHS Digital, n.d.-f, 2018d). The existence of the right balance between the protection of confidential data and the use and sharing of them is crucial (Department of Health, 2013a; Department of Health, UK, 2013).

When conducted in an appropriate way, the reuse of patient data produced in the healthcare sector can provide benefits for patients and the whole population (Custers & Ursic, 2016) since these data can be used for example to identify, prevent and manage risk factors for health conditions (Hays & Daker-White, 2015). Nonetheless, the use of patient data in some contexts beyond the healthcare sector can raise critical privacy and power issues (Harwich & Laycock, 2018; Powles & Hodson, 2017). Power relations are important to consider because they shape and mediate ideas, and social relations play a critical role in the mediation of both ideas and objects (Howell, 2013). In the case of health data, power relations may come into play when determining what knowledge claims are valued and devalued, how data are produced and used, who can have access to them and under which circumstances.

In recent years, a number of the data sharing initiatives between the UK's National Health Service and external parties have been particularly controversial. For example, in 2014, the Care.data initiative (now cancelled) was proposed with the objective of sharing patient data generated in primary care settings with external organisations without informing patients what data would be shared and with whom. This initiative generated a lot of confusion because of the lack of information given to the population about what Care.data would involve, what data would be available to whom and in what form (Kirby, 2014). Citizens raised concerns regarding lack of transparency, and insufficient respect for confidentiality and privacy (Sterckx et al., 2016). On Twitter this initiative was described by citizens as improper in terms of scope, privacy, transparency and management (Hays & Daker-White, 2015).

In November 2015 health records of NHS patients held by the Royal Free London Trust were transferred, without explicit consent from patients, with Google DeepMind (Powles & Hodson, 2017). Around the same time, personal data from NHS patients were shared with the Home Office to trace individuals tagged as 'potential immigration offenders' (Iacobucci, 2018). More recently, it was revealed that international pharmaceutical companies such as Bristol Myers and Eli Lilly have obtained access to NHS patient data. A number of the partnerships between

the NHS and external parties that have enabled the flow of patient data from the healthcare sector to technology and pharmaceutical companies have appeared not to have taken into account patient privacy and consent. These partnerships have also been handled with a large degree of opacity and failed to openly communicate to patients how their data is being used. The details of the sharing agreements signed as a result of these partnerships have not been voluntarily disclosed by the parties involved, rather they have been revealed to the public through journalistic investigations (Powles & Hodson, 2017).

There are therefore big questions to be asked about the movement of patient data generated in the NHS towards external sites to be reused for purposes other than the direct care of patients. The circulation of health data towards external sites to be used for secondary purposes has been studied extensively. This body of literature often has a techno-managerial orientation and pays great attention to identifying data quality issues that create barriers for data to flow and how to address such issues; generally in balance with privacy concerns. Patients' and the public's attitudes concerning the circulation of health data beyond the healthcare sector for secondary purposes have also received significant attention. However, less research has been conducted focusing on the underlying sociocultural and material factors (e.g. infrastructure and material conditions of production) that are shaping the circulation of digital data between different groups of social actors and sectors (Bates, 2017). This is important because power dynamics embedded in these socio-material processes play a significant role in shaping data flows and data practices.

In an effort to address such concerns, in recent times a number of scholars in Critical Data Studies and related fields have been paying increased attention to such aspects of the nature of digital data flows, particularly the ways in which different social, cultural, and political factors influence how data move in different contexts and the potential social implications of emergent digital data movement (Bates et al., 2016; Beer, 2013b; Merricks, 2017). However, little work in this field has looked at health data flows. Within this field Aula's (2019) work is one of the few studies that have paid attention to data flows in the healthcare sector. However, he pays attention to the infrastructural and legal reform of secondary uses of data in Finland.

In the context of increasing concern about privacy and power around health data, this research adopts a Critical Data Studies (CDS) approach which is characterised by paying attention to such social, cultural, political and ethical issues related to the ways in which power dynamics play out in digital data production, use and movement (Iliadis & Russo, 2016a; Kitchin & Lauriault, 2014). The research presented is an in-depth qualitative study of health "data

journeys" (Bates, Lin, & Goodale, 2016) that examines the socio-material drivers and frictions that shape patient data flows as they move from NHS providers to various academic institutions to be reused for research purposes..

1.2 Research aim, objectives and questions

1.2.1 Research aim

The aim of this research is to gain understanding of how sociocultural values and norms interact with existing and emergent material conditions to shape"data journeys" (Bates et al., 2016) of patient data in the UK health sector, with a specific focus on reuse of patient data for research purposes. It will examine where particular types of personal health data flow, and the socio-material drivers and frictions that represent the politics of these data flows. This research will also explore the potentials of "Data Journeys" (Bates et al., 2016) as an approach to investigate the movement of personal health data produced in the healthcare sector.

1.2.2 Research questions

Main Research Question

- **RQ1** How do sociocultural values interact with existing and emergent material conditions to generate drivers and frictions that work together to produce the material forms of data journeys in the context of reuse of NHS patients' personal data for research purposes?

Sub-Research Questions

In order to answer the main research question, the following sub-questions are posed:

- **RQ2** What material forms do the 'journeys' of NHS patients' personal data take between different sites of practice, from their initial generation though to reuse for research purposes in different contexts?
- **RQ3** In what ways do sociocultural values interact with existing and emergent material conditions to act as drivers to move data between different sites?
- **RQ4** In what ways do sociocultural values interact with existing and emergent material conditions to act as frictions on efforts to move data between different sites?

1.2.3 Research objectives

In order to achieve the research aim and answer the research questions, five objectives are established:

1. To investigate the different ways in which the movement of data has been studied in the past and to explore the contributions and limitations of such studies.
2. To examine selected data journeys of patient data generated in the UK healthcare sector that are reused in universities for research purposes.
3. To gain insights into the sociocultural values and norms within university-based research groups and how they act as factors driving or constraining the movement of patient data towards university-based research groups.
4. To gain insights into the material conditions that act as factors driving or constraining the movement of patient data towards university-based research groups.
5. To reflect on the potential challenges that can be encountered when applying the Data Journeys approach to explore patient data flows and how might this approach be adapted for future use in similar contexts.

1.2.4 Expected outcomes

The proposed research intends to achieve the following outcomes:

- Generate new knowledge on how and why socio-material factors shape personal data flows within the healthcare sector.
- Provide insights for practitioners to give them a better understanding of their own – and others' – motivations and assumptions when trying to influence the circulation of data.

1.3 Thesis structure

This thesis is organised into nine chapters as follows:

Chapter 1. Introduction. This chapter provides an introduction to the thesis by presenting the background to research topic, aim of the research, research questions and objectives, and expected outcomes.

Chapter 2. Literature Review. This chapter provides an overview of existing literature. This chapter has two objectives; first, it presents a review of the different ways in which the movement of data has been studied in the past. Second, it explains the approach adopted to conduct this study.

Chapter 3. Methodology. This chapter discusses the methodology used in this research. The rationale that supports the selection of procedures and techniques to conduct the research is presented. First, the philosophical position of this research is explained. Afterwards, the rationale for the selection of the research methods used in this project is provided. Later in this chapter, the data collection strategy and analysis techniques that were applied in this research are presented. Finally, research quality and ethical considerations and issues related to this study are addressed.

Chapters 4, 5, 6 and 7. Findings. Chapter 4 presents detailed descriptions and diagrammatic representations of five selected data journeys of patient data that were followed as part of this study. Chapter 5 presents the socio-material factors that have driven the movement of patient data towards university-based research projects or groups. Chapter 6 presents the socio-material factors that have generated frictions for the movement of patient data towards university-based research projects or groups. Chapter 7 discusses the role of the support of patients and the public in strategies adopted by university-based researchers to smooth data frictions.

Chapter 8. Discussion. In this chapter the findings are consolidated and interpreted in the light of the literature review and integrating two strands of thinking: Data Journeys (Bates et al., 2016) and Data Valences (Fiore-Gartland & Neff, 2015).

Chapter 9. Conclusion. This last chapter revisits the research questions of the study and discusses how these questions were answered. It also explains the empirical, theoretical, and methodological contributions of this study. The chapter finally presents recommendations to university-based researchers and suggestions for future research.

Chapter 2. Literature Review

2.1 Introduction

This chapter has two objectives; first, it presents a review of the different ways in which the movement of data has been studied in the past and explores the contributions and limitations of such studies. Second, it explains the approach adopted to conduct this study. In the first section I discuss the concept of data. In section 2.3review how data circulation has been explored in Information Studies and related fields, paying particular attention to the approaches that scholars from this field have adopted, their objectives and main contributions. In section 2.4 I discuss how health data flows have been explored, highlighting the main contributions and limitations of this body of literature. As stated before, this research explores the movement of personal health data generated in the UK healthcare sector, adopting a Critical Data Studies approach. Thus, in section 2.5, I describe this approach, explaining how this emergent field came to existence, highlighting its key characteristics. In this section I also provide some examples of empirical research in CDS and related fields. In section 2.6 of this chapter, I introduce the approach I am using in this research to explore the movement of data. This section explains the data journeys methodology (Bates et al., 2016), describes the CDS conceptual framework it sits within and explains the reason for adopting it in my study.. Finally section 2.7 provides an extended discussion about the concept of data valences (Fiore-Gartland & Neff, 2015) and explains the reasons for using it in this research.

2.2 Data and their nature

Data is considered a highly valued resource in the modern world and its nature has been analysed in many different contexts (Borgman, 2015; Kitchin, 2014b). Common assumptions about data that suggest data are neutral, objective, and pre-analytic in nature or that define them as only the 'building blocks' or the 'raw material' to create information and knowledge have been labelled as limited (Lauriault, 2017). These ideas have been contested by a number of scholars who understand data in a much more complex way (Borgman, 2015b; Gitelman & Jackson, 2013; Kitchin, 2014d). Scholars who reject the idea that data are independent objects highlight that data do not emerge from nowhere and their generation is not spontaneous (Kitchin & Lauriault, 2014). Rather, their production is a result of human action (Edwards, 2013; Iyamu & Mgudlwa, 2018). Therefore, data cannot be understood as neutral and objective because they are always infused with the assumptions and politics of their creators and

calculators (Merricks, 2017). Data cannot exist independently of the ideas, tools, contexts, theories, and practices that contribute or intervene to their generation, processing, and analysis (Kitchin, 2014a); data are not simple evidence of phenomena, they are phenomena in and of themselves (Wilson, 2015). The ways in which data are processed and analysed changes over time and are influenced by several factors, such as changes within organisations, legislative developments in data management and data protection, technological advances, the adoption of new data sorting and analysis strategies (Kitchin, 2014a). In Gitelman and Jackson's (2013, p. 2) words, they are "'always already 'cooked' and never entirely 'raw'." In this sense, it can be said that the world is not just reflected in data (Ribes & Jackson, 2013) but data perform active work in the world (Kitchin & Lauriault, 2014, p. 11).

This research embraces the notion that data are not neutral, objective, and pre-analytic representations of the world (Gitelman & Jackson, 2013). Rather, they are generated in complex contexts that are always influenced by systems of thought, norms and regulations, organisational processes, technical instruments, and economic forces (Kitchin, 2014b; Kitchin & Lauriault, 2014). It is also recognised that regardless of how large in volume they are, data will always be limited and incomplete representations of the world (boyd & Crawford, 2012). Taking this logic forward it is possible also to understand the flows of data (the journeys data take) to be shaped by similar social contexts. In the following section I will discuss how data flows have been imagined in Information Studies and related fields, then in section 2.4 I will discuss how data flows have been explored in the past. As mentioned earlier, later in this chapter I will discuss the theoretical orientation adopted in this research and provide an explanation about the reasons for adopting the data journeys approach (Bates et al., 2016) and integrating the concept of data valences (Fiore-Gartland & Neff, 2015).

2.3 Researching the movement of data in the Information Studies and related fields

In the Information Studies field, the movement of data has been extensively explored. In recent years a number of in-depth qualitative case studies have examined data flows and data reuse practices in different contexts (Borgman et al., 2018; Sands et al., 2012; Yakel et al., 2019).

For example, Sands et al. (2012) investigated data reuse practices of astronomers, paying particular attention to the circulation of data feeding into and out of research outputs of astronomers with the objective of understanding the people and infrastructures that are involved in developing, sustaining, and curating large sky surveys. Borgman et al. (2018) conducted a study to investigate the roles and relationships of data contributors, data users, and

data curators. This work paid attention to the traits and capabilities of knowledge infrastructures that support the exchange of data and the mediating roles performed by archives as organisations and by archivists as collaborators with contributors and consumers. More recently, Yakel and colleagues (2019) explored the 'trajectory of data' in order to understand how the data production, sharing, curation and reuse practices of zooarcheologists positively or negatively affect different stages of the life-cycle of data. In contrast to Sands et al. and Borgman et al. who paid attention to both people and infrastructures, the research of Yakel et al. focused mainly on the role of people.

In related sub-fields of Science and Technology Studies, Sabina Leonelli (2016) has had a long interest in the movement of data across infrastructures. Similar to Bates et al. (2016) Leonelli (2016, p. 39) utilises *data journeys* as a notion to refer to the movement of scientific data from the initial site in which they are generated, to different sites within or beyond the research field in which they are produced. However, the metaphor of "data journeys" has a different interpretation, focus and disciplinary approach in Leonelli's work. Leonelli's work (2016) has a strong emphasis in techniques and procedures involved in the handling of data. She is interested mainly in the technical conditions that are necessary for data to move in different research situations, specifically in the field of biology, and in the diverse ways in which the value of data can be determined. She pays particular attention to processes of packaging data that "prepare data to travel". These procedures involve the classification, selection, formatting and standardisation of data and the creation and utilisation of methods to analyse, retrieve, visualise and control the quality of data. A comparative table presenting the key differences between Leonelli's (2016) and Bates et al.'s (2016) approaches can be found in appendix 1.

Other researchers within the STS field have shown interest in the movement of data. Edwards (2013) developed the concept of "data friction" to explain what happens at the different interfaces when climate data moves between people, substrates, organisations or machines. From his perspective, flow of data is not something that just happens without any constraint, rather he argues that every movement of data across different interfaces comes at a cost in time, energy and human attention (Edwards, 2013). Friction, a term that emerged in the field of physics, is understood as resistance; friction takes places between objects or surfaces at different interfaces (Edwards, 2013). As defined by Edwards, data friction is "the cost in time, energy and attention required to collect, check, store, move, receive and access data" (Edwards, 2013, p. 84). As data friction hinders the movement of data whenever they travel, Edwards

perceives data friction as an obstacle to scientific advancement, and suggests that efforts must be made to overcome data friction (Edwards, 2013).

Approaches such as the ones adopted by Borgman, Sands and colleagues and Edwards do not pay attention to matters that are central within the CDS approach, such as ethical, political, power, and social issues related to data. Their work may sometimes touch on social, cultural, political or other CDS-related aspects, but their mentions mainly remain at a surface level. They recognise for example the social nature of data and the fact that they are not neutral (Borgman, 2015b); however they do not explore in depth CDS-related aspects.

Other researchers in the Information Studies field have shown interest in understanding similar aspects of the movement of data, with a focus on the development of best practices to overcoming barriers to data flows in a variety of contexts; a large number of these information studies work have tended to pay attention to managerial aspects of data flows. For example, the qualitative study of Dallmeire-Tiessen et al. (2014) paid particular attention to the drivers, barriers and enablers for scientific data to be shared and reused. Findings of this work suggest that within e-science the sharing of research data is both a challenge and an opportunity, and that in order to maximise the reuse of data, barriers for data sharing must be overcome. Examples of the barriers for data sharing identified by Dallmeire-Tiessen and colleagues include availability of a sustainable preservation infrastructure, data usability, financial constraints and legislation. They developed a conceptual model with the objective of helping researchers, policymakers, funders and service providers to identify potential barriers they might encounter in data sharing processes, so that they can design strategies to overcome these barriers and maximise the reuse of data.

Bote and Termens (2019) are other Information Studies scholars who have paid attention to the movement of scientific data. Drawing from a review of literature, they analysed the ethical and technical challenges for data generated in European research centres, universities and public organisations to flow outside of their creator organisations to be reused for research purposes. One of their key arguments is that in order to be able to fully profit from data sharing and reuse, it is necessary to overcome a number of barriers (Bote & Termens, 2019); the barriers identified in their work include lack of accuracy in documenting data practices and privacy and data protection policies. This work highlights that sharing of scientific data can be beneficial for natural disaster prevention, climate change mitigation, and mobility enhancement in urban environments; however it does not address what the implications of data sharing are and what could be potential negative consequences of promoting smooth flows of data. Bote and

Tremens (2019) proposed a set of 'best practices' to overcome the aforementioned data sharing barriers which include, ensuring that accurate metadata is provided to help other researchers to make sense of data, sharing data using specialised repositories, and delivering training to researchers on data creation process.

In 2016, De Roo and colleagues (2016), conducted a study aimed at exploring the flows of administrative, spatial, and scientific data across data infrastructures within an archaeological organisation. Their objective was to provide insights about how data is managed in Flemish archaeological processes and how this management process could be improved. They paid attention to the type of data that is produced, stored and used in archaeological processes and the stakeholders involved in these data practices. Adopting a document-analytical and a practical approach to address their research questions, this research's ultimate goal was to propose a series of recommendations to enhance the management of data flows. Their findings highlighted that metadata registration was a key issue in the Flemish archaeological field as it was common to find transcription errors and data inconsistency in records. They also showed that since data is frequently shared digitally, records in paper format may not be easily shared. In order to enhance the information management, De Roo and colleagues proposed the development of an archaeology-specific data infrastructure aimed at bringing a cost and time-efficient solution for fostering the flow of information. Other researchers in the Information Studies field have paid attention to the flow of data within organisations. For example, Sugahara (2019) examines how to make data flows within organisations more effective. They highlight that it is key to have effective information flows within an organisation in order to help the development of individual and collective knowledge. Sugahara argues that since relations between individuals within an organisation can foster the movement of data, reinforcing connections between people can be helpful to smooth data flows. The majority of studies that have investigated the movement of data within organisations have tended to explore the movement of data with the intention of finding ways or making data flows more smooth to make the processes within organisations more efficient, or how to make data flow without constraints.

As discussed above, in the Information Studies discipline, a large number of studies have focused on the processes that take place inside information systems with a focus on the management of data flows. However little attention has been paid to the ways in which sociocultural factors interact with existing and emergent material conditions to shape these flows.

2.4 Research on health data flows

Data flows have received significant attention from the health informatics field. In this body of literature, several studies have focused their attention on specifically exploring the flows of data beyond the healthcare sector for secondary purposes with a focus on research reuse. This literature has been dedicated to exploring mainly the following aspects: 1) Data quality and data management related factors that make the flow of health data for secondary purposes more difficult, and 2) attitudes of different stakeholders concerning the reuse of health data, focusing mainly on patient and public perspectives, but also in some cases exploring the views of actors such as healthcare professionals and policymakers. In this field other aspects related to the circulation of data for research purposes have also been explored, however in a less intensive way than these two aspects. Examples of other topics that have been explored include information governance, and legal and ethical aspects that can pose barriers for the reuse of patient data. These studies have been mainly concerned with identifying such barriers and in some cases also developing approaches to overcome them. In a similar way to the Information Studies literature reviewed above, this body of research is different to the Critical Data Studies approach adopted in this study in the sense that their main interest has not been paying attention to ethical, political, power, and social issues related to data and data flows (Iliadis & Russo, 2016b). An overview of this body of literature is presented in in this section.

2.4.1 The impact of health data quality on data flows

In recent years several studies about the reuse of medical data have paid attention to the challenges and barriers to reusing health data for purposes other than providing the right care to patients (Edmondson & Reimer, 2020; Kim et al., 2019; Meystre et al., 2017; Schlegel et al., 2017; van Velthoven et al., 2016). While this body of literature has centred its attention in understanding the barriers for the reuse of data, these also impact the circulation of data. According to the findings of three literature reviews, exploring data quality issues has been a priority for researchers in recent times (Edmondson & Reimer, 2020; Meystre et al., 2017; Schlegel et al., 2017). Meystre and colleagues (2017) conducted an extensive review with the objective of providing an overview of the studies about the reuse of medical data between 2005 and early 2016. Several studies have paid attention to data quality issues as they are considered key barriers for reusing medical data. Meystre et al. (2017) highlighted that medical datasets' data quality has been frequently referred to as insufficient or problematic. Findings of a number of studies examined as part of this review concluded that datasets are often incomplete (Thiru

et al., 2003; Weiskopf et al., 2013), patient records are fragmented, and data is inaccurately recorded (Kahn et al., 2012; Weiskopf & Weng, 2013). Meyste and colleagues' analysis pointed out that in the studied period, several studies have been conducted to design approaches for assessing data quality: examples of studies of this nature are the works of Thiru and colleagues (2003) and Weiskopf et al. (2013).

Schlegel et. al. (2017) review studied research published in the databases Medline and Web of Science, with one of their objectives being to summarise research and emerging trends in the area of secondary uses of healthcare data. A total of 14 journal articles and paper proceedings published in 2016 in Medical Informatics were selected based on relevance and perceived impact. In a similar fashion to what was reported by Meyste et al., Schlegel and colleagues' key finding was that data quality issues (e.g. missing data and lack of consistency) act as barriers for reusing health data. An interest in addressing data quality issues had motivated some researchers to conduct research aimed at identifying and designing methods or technical solutions to address such issues (Schlegel et al., 2017). Examples of studies of this nature highlighted by Shlegel et al. are the work of Sahoo et al. (2016) and Sáez et al.(2016). The former implemented an informatics platform to identify heterogeneity issues and developed a strategy to solve them. In contrast, the latter designed a statistics-based assessment method to identify data quality issues.

More recently, Edmonson and Reimer (2020) conducted a systematic review to identify issues concerning the secondary use of patient data from electronic medical records. These authors analysed sixty articles published up to June 2018 in Medline and CINAHL. These authors, as Myestre and colleagues (2017) and Schlegel et al. (2017) reported, found that one of the main barriers for the reuse of patient data is that data quality is insufficient. This impacts the flow of data because if data lacks in quality, additional work might be needed to clean data before it can be reused. It might also be the case that the circulation is totally stopped if potential reusers abandon their intentions of working with these data if they do not have the resources to address the data quality issues in order to be able to reuse these data.

Researchers have also shown interest in conducting comparative explorations about the barriers to reusing medical data across different countries. One example of a study of this nature is the work of Velthoven and colleagues (2016). As part of this study, they conducted a qualitative exploration to identify the challenges to extract data from electronic medical records (EMR) for research purposes across seven high- and middle-income countries: Italy, Saudi Arabia, Republic of Korea, Taiwan, United Arab Emirates, Brazil and South Africa. A key finding of

this study was that insufficient data quality is a significant barrier that act as frictions for the extraction of medical data to be reused. In countries such as Italy, Republic of Korea, Saudi Arabia, Taiwan and the United Arab Emirates, with high fill rates of electronic medical records and comprehensive data availability, it is more feasible to reuse patient data extracted from electronic medical records with fewer constraints. On the other hand, in countries where EMR data quality is modest such as Brazil and South Africa, it is more challenging to extract data for research purposes. This comparative analysis sheds light on the issues related to filling in patient records and how this impacts on data quality and therefore data flows. Velthoven et al.'s (2016) key findings concerning data quality are in line with that previously reported in other studies: quality issues such as missing data and inaccurate data are key barriers towards the reuse of medical data.

As observed, quality issues have been identified as significant barriers to the circulation of data from sites of production to sites of reuse. In response to this, scholars have undertaken efforts to developing quality evaluation strategies to investigate whether datasets generated in the healthcare sector are suitable to be reused for given research purposes. Weiskopf et al. (2013), for example, conducted a study to examine different methodologies used to evaluate the quality of data within electronic health records. They identified that the following dimensions are commonly assessed to determine the quality of data: "completeness, correctness, concordance, plausibility, and currency" (Weiskopf & Weng, 2013, p. 144). According to their findings, completeness seemed to be the most important dimension for researchers and the most commonly assessed. They also detected different types of methods to assess the quality of data, such as gold standards comparison, data source consistency, and validity checks (Nicole Gray Weiskopf & Weng, 2013, p. 144). This study pointed out that whereas different methods can be adopted to evaluate the quality of data, one of the key issues with them is that in general terms they lack in consistency.

In a more recent study, Kim and colleagues (2019) also paid attention to barriers to the use of data movement from the healthcare sector for secondary purposes. Their work centres its attention on exploring the application of computer science and artificial intelligence to overcome critical challenges for reusing patient data derived from data quality issues in the US context. These authors coincide with the scholars cited above, saying that data quality is one of the main challenges for reusing patient data. These authors agreed that data recorded in electronic health records are often incomplete, messy, inaccurate, and heterogeneous. They also pointed out that despite gradual improvements in the US context, they are still far from a

"Complete EHR" (Kim et al., 2019, p. 354). Findings of this study suggest that the application of Natural Language Processing methods to mine and extract data from health records have become increasingly popular in recent years. This is because it is believed that they enable the exaction of quality data in a less time- and cost-intensive way.

As stated above, a large number of studies have focused on identifying data quality issues present in health datasets and how to overcome or address such issues. A study that also paid attention to data quality issues but with another purpose is the work of Martin-Sanchez et al. (2017). Their work explored the different risks of bias that can emerge when data quality issues are identified within datasets, and highlighted that researchers should be aware and make efforts to avoid them. They pointed out that since patient data is recorded for providing direct care to patients, it cannot be expected that such datasets offer a complete and accurate medical picture, and certainly not a holistic representation of the health population under examination. Martin-Sanchez and colleagues have stressed that a number sources of bias can have an impact on research using routinely collected health data. The work of Martin-Sanchez et al. (2017), in contrast to the majority of other studies on health data flows, does not encourage potential reusers of patient data to develop approaches to improve the quality of data to allow data to flow with fewer constraints. Instead, it highlights the potential risks of working with this type of data.

To summarise, data quality is one of the most explored aspects in the literature regarding the secondary uses of data for research purposes. This is because data quality issues have been considered key barriers for allowing data flow beyond the healthcare sector for research purposes. The studies conducted have contributed to identifying that datasets produced in the healthcare sector are often incomplete, fragmented and inaccurate. Therefore, it cannot be assumed "that such data provide a full or accurate clinical picture, let alone a full description of the health of the population under study" (Martin-Sanchez et al., 2017, p. 30). This body of literature has also been helpful to understand that these limitations have motivated researchers to investigate how data quality issues can be addressed. Such research has mainly been concerned with finding strategies and approaches to remove the barriers (e.g. quality issues) that prevent data moving from the healthcare sector to external sites to be reused for research purposes. If the quality of data is considered low, researchers might not want to use them in their projects. It also can be that more work might be required to get data ready to move from the site of data production to sites of reuse, which might slow down or completely stop the movement of data.

2.4.2 The impact of patients and other stakeholder views on data flows

The attitudes of patients and the public concerning the flow of health data generated in the healthcare sector towards external organisations for reuse purposes have also been largely explored in the literature. Research in this area has been mainly focusing on exploring levels of support of data reuse outside the healthcare sector, conditions to support the reuse of data, awareness of and main concerns concerning data reuse practices, and perceptions concerning different stakeholders accessing data for research purposes. This body of research 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, de St Jorre, et al., 2016) or the design of strategies to foster patient and public trust in organisations conducting research with patient data (Kalkman et al., 2019).

2.4.2.1 Public and patient views about the reuse of patient data for research purposes

This section highlights key findings in the literature regarding the perceptions of patients and members of the public about the use of patient data for research purposes. While this section mainly highlights findings from studies conducted focused solely in the reuse of patient data, where appropriate it also highlights key findings of studies with a wider coverage i.e. those that pay attention to data practices of other actors and sectors. While the main focus of this section is to discuss the views of people about the reuse of patient data for research purposes reported in the literature, to put this in context it is important to also pay attention to the attitudes towards reuse of data for other purposes.

In general, members of the public (Aitken, de St Jorre, et al., 2016; Skovgaard et al., 2019) and patients (Grande et al., 2013) have little or limited knowledge about how patient data is used for research outside the healthcare sector. For example, a study conducted in the UK in 2015 reported that only 18% of participants are aware of the ways in which academic researchers use health data (Wellcome Trust, 2016). While understanding of data uses is limited, several studies reported that the public and patients had expressed interest in obtaining more comprehensive detail concerning the ways in which data are reused by researchers and the measures adopted to prevent abuses in the use of data (Aitken, Cunningham-Burley, et al., 2016; Grande et al., 2013).

Public views on the acceptability of sharing patient data is not the same for different purposes. As reported in the literature, the majority of people support sharing patient data for the direct

care of patients (Understanding Patient Data, 2018). For example, a study conducted in the UK in 2015 found that 96% of participants support this type of data use. On the other hand, only 77% supported sharing patient data for research purposes (Wellcome, 2015). A similar figure was reported by another study conducted in 2010, also in the UK, in this case, 74% of respondents expressed support for sharing patient data to conduct medical research (New Economics Foundation, 2010). However, much lower numbers were reported concerning the support of use of data for research purposes in a survey conducted more recently (Hartman et al., 2020). In this study, only around half of the participants (52%) expressed support for this type of use, although it is worth mentioning that the question posed was different. Participants were asked whether they would support sharing personal data for research in the public benefit. One factor that could have influenced the lower rate of acceptance is that, participants were asked about personal data, but not specifically about personal patient data. As will be discussed in more detail below, a key factor that leads people to support the sharing of their patient data for research purposes is the potential of helping future patients or people with similar health conditions (Kalkman et al., 2019). Other prevalent reasons for support for sharing patient data expressed by patients are the value in answering "crucial" research questions and a desire to contribute to medical advancements (Skovgaard et al., 2019).

While the levels of acceptance regarding the reuse of patient data for research purposes reported in the literature are varied, the findings about the perceptions of using data for marketing or insurance purposes are very similar. Studies in the UK are consistent in pointing out that members of the public show resistance to share their patient data to be used for either of these purposes (Wellcome Trust, 2016). Regarding the use of patient data for profit, contrasting views have been reported. A study conducted in 2017 found that using patient data for profit is acceptable for people only if it does not undermine public benefit (Chico et al., 2019). In contrast, Hartman and colleagues (2020) reported that 78.3% of respondents were not in favour of using personal data for profit. Similarly, a more recent survey conducted in 2020 found that people do not support the use of personal data for profit (Kennedy, et al., 2021).

While the above studies are UK focused, it is worth mentioning that positive views about the reuse of patient data for research purposes have been reported in studies conducted in different parts of the world. For example, Aitken and colleagues (2016) pointed out widespread support from the public to reuse health data for research purposes across the different countries in Europe and North America. Similar findings have been reported in other places such as Ontario, Canada (Perera et al., 2011), the United States of America (Grande et al., 2013), and

Denmark (Holm et al., 2020). The contrasting findings reported in the literature suggest that as Kennedy and colleagues argue, the way in which research is carried out makes a difference to what it finds (2021). Factors that play an important role in shaping findings and the claims authors make include the research methods, types of questions asked to participants, they way in which findings are interpreted and presented, and the disciplinary background and the political orientation of the research team.

Three extensive literature reviews conducted in recent times (Aitken, de St Jorre, et al., 2016; Kalkman et al., 2019; Skovgaard et al., 2019), as well as several qualitative studies, are consistent in highlighting that many patients and members of the public supports the reuse of patient data for research purposes, nevertheless they also stress that this support is not unconditional and some of them discuss the key factors that lead people to support sharing.

2.4.2.2 Conditions to support reuse of patient data

As mentioned above, the majority of the studies in the field of health data point out that while both public and patients support the reuse of patient data for research purposes, this support is not unconditional (Aitken, de St Jorre, et al., 2016; Kalkman et al., 2019; Perera et al., 2011; Skovgaard et al., 2019; Tully et al., 2018). The key condition reported in the literature for support or acceptance from patients and the public in general, is that data are used for the public benefit, or in other words "for the greater good" or “in the public interest” (Aitken, de St Jorre, et al., 2016; Kalkman et al., 2019; Skovgaard et al., 2019; Tully et al., 2018).

Confidentiality also appeared to be of great importance for obtaining public and patient support. According to the findings of a number of studies (Aitken, de St Jorre, et al., 2016; Kalkman et al., 2019; Perera et al., 2011), people have expressed that they are willing to support the reuse of patient data for research purposes only if confidentiality can be maintained and data shared anonymously. A contrasting finding concerning this aspect emerged from the study conducted by Holm et al. (2020). They explored the views of Danish patients' about the secondary uses of patient data. Participants of this study did not believe anonymisation is compulsory for them to support research using patient data. According to the authors this is perhaps because in Denmark the management of health data is centralised, which means that identifiable patient data are gathered in centrally-managed databases allowing the linkage of data produced in the healthcare sectors with external datasets via a unique identifier.

The use of patient data for commercial gain according to some studies is considered acceptable only if it also produces health benefits (Chico et al., 2019; Connected Health Cities, 2017), as

people consider that commercial gain should be secondary to public benefit. In contrast, people consider it unacceptable to use patient data only for profit, even if safeguarding measures are in place (Connected Health Cities, 2017). While studies focused on views about the use of patient data point out that people do not always oppose the reuse of patient data and that they tend to accept it as long as it does not undermine the public benefit, studies that focus on public perceptions regarding the use of personal data in general, not specifically patient data, have reported contrasting results. For example a large number of respondents (78.3%) of a survey were against the use of personal data for profit by commercial companies (Hartman et al., 2020). Similarly, a more recent study involving some of the same researchers found out that people do not support the use of personal data for profit generation (Kennedy, Taylor, Oman, Bates, Medina-Perea, et al., 2021). These differences in the responses of participants suggest that attitudes depend on a number of factors, not only on who is accessing data or for what purposes, but also the potential for public benefit and helping others. As Kennedy and colleagues put it, context matters (2021).

As noted, several studies have reported that the public prefers data used for research to be anonymous. However, people do not have a clear understanding of some of the terms used to talk about levels of anonymisation, for example there is confusion regarding the difference between potentially identifiable data and anonymised data. Therefore, it is difficult to fully understand what members of the public mean when they express that they prefer data to be anonymised (Aitken, de St Jorre, et al., 2016). Whereas assurance of anonymisation seemed to be essential for the public, some studies unveiled more subtle views, highlighting for example that when people are allowed a deeper reflection on the implications of anonymisation, they frequently recognise that the use of anonymised data does not guarantee confidentiality, and therefore become more concerned (Aitken et al., 2016; Wellcome Trust, 2016). Similarly, Tully and colleagues (2018) reported that, whereas some people tend to express less concern about the reuse of patient data as they know more about its potential benefits and risks, not everyone responds in this way. Some people can become much more sceptical rather than less (Tully et al., 2018). Their research suggests that different individuals may well receive the same information but reach different conclusions, perhaps because of differences in values. Kennedy and colleagues (2021) have also pointed out that people who are more knowledgeable about data uses are more likely to have negative attitudes towards them.

2.4.2.3 Key public concerns regarding patient data reuse practices

The use of patient data raises a number of concerns, and these can vary across different demographic groups. This section will explore the concerns commonly shared by people before paying attention to concerns that are specific to certain groups.

Key general concerns identified in the literature include data security and misuse, and the potential to use data for harm. Regarding security, people have expressed concern about hacking, data leakage, data loss and unauthorised access (Stockdale et al., 2018). An extensive literature review reported that a number of studies found that patients and members of the pubic expressed concern about data falling into the wrong hands. This is a matter of great concern because people believe that this could lead to misuse of data (Aitken, de St Jorre, et al., 2016). People are concerned that data could be used for harm (Stockdale et al., 2018). To be more specific, people are concerned that outcomes of research could be used for: 1) allowing or perpetuateing mass surveillance; 2) enabling the tagging of vulnerable societal groups that could result in discriminatory treatment or stigma; and, 3) conducting research based on analysing large datasets to inform the design of policies for the whole population instead than considering particular circumstances and requirements.

Little research has been conducted to understand the perceptions of groups with specific health conditions regarding the use of patient data, however one important finding from this literature is that people with specific health conditions tend to be more supportive of certain uses of patient data that bring a direct benefit to people suffering the same condition as them. For example a study exploring the views of people with asthma reported that 94% of participants felt comfortable with the idea of sharing of anonymised data with an analytics company to support the development of tools for people at risk of suffering an asthma exacerbation (Asthma UK, 2018). Similarly, Cancer Research UK and Macmillan Cancer Support reported that 85% of people with cancer support the reuse of cancer data to improve services and medical treatments, while only 72% of the UK general public supported this (2016). These findings contrast with the findings from Chico and colleagues (2019), which reported that only 45% of the general public support sharing patient data with commercial companies.

A study that demonstrates difference in attitudes across different groups to data uses is the recent Living with Data Survey conducted by Kennedy and colleagues (2021). According findings of this work certain demographic groups such as LGBTQ+ people, people of colour and younger people reported to be more concerned about data uses than other groups. The study

also highlighted that although higher levels of concern were reported by some groups, the difference were not large and while there are certain groups that are most likely to be harmed by data practices they are not always the ones that are most uncomfortable with them. Overall, they found that factors such as knowledge and attitudes play a key role in shaping perceptions.

2.4.2.4 Trust in data sharing and reuse

According to findings from the literature, people have different levels of trust in different sectors or institutions. A number of studies pointed out that people are inclined to trust more in public sector than in private actors. A number of studies have reported that the highest levels of trust correspond to GP surgeries and the NHS (Ford et al., 2017; Given et al., 2017; Kennedy, et al., 2021; Wellcome Trust, 2016), followed by researchers in academic institutions. The public and patients tended to express positive attitudes about sharing patient data with researchers (Perera et al., 2011). A number of studies identified that university-based research teams are often perceived as trustworthy (Aitken, Cunningham-Burley, et al., 2016). In contrast, pharmaceutical or other commercial companies are less trusted. For example, according a number of studies people have a strong tendency to perceive as negative pharmaceutical actors having access to health data (Perera et al., 2011; Skovgaard et al., 2019).

Something important to keep in mind when thinking about trust is that context is really important, as findings of the recent Living with Data survey shows, trust in sectors or institutions strongly influences the trust of people in the same sectors' and institutions' uses of data (Kennedy, et al., 2021).

Research also suggests that scandals have the potential to undermine trust in organisations. For example, a survey conducted in the UK in recent years found out that the confidence of participants in the ability of the NHS to handle data was negatively impacted after becoming aware of the Wannacry hacking scandal (Healthwatch England, 2018). The Care.data scandal also generated a negative impact. When this initiative was launched citizens expressed that it was improper in terms of scope, privacy, transparency and management (Hays & Daker-White, 2015) and around 1.5 million people opted-out of their data being used outside of direct care (Vezyridis & Timmons, 2017). While scandals can undermine trust, research has also highlighted that despite a number of recent scandals related to controversial data sharing initiatives from the NHS, the trust in this organisation remains high (Neves et al., 2019; Skovgaard et al., 2019).

According to Aitken and colleagues, it is necessary to raise awareness among the public concerning reuse practices of patient data and provide good opportunities for them to engage in meaningful discussion about such practices. These actions are considered vital by many research teams to legitimise research projects that reuse patient data and to minimise the risk of generating controversies. According to these authors researchers increase their possibilities of being perceived as trustworthy if their engagements with the public are open and transparent, and if they invite the public for a dialogue and go beyond merely sharing information with them or treating them just as listeners (Aitken, Cunningham-Burley, et al., 2016, p. 29).

2.4.2.5 Other health stakeholders' views on patient data reuse for research

Besides patients' and the public's views, other stakeholders' perceptions have been explored in the literature; however, in a less intensive way. Studies have been conducted focusing on the attitudes of healthcare professionals and policymakers. Two studies conducted in the UK that are relevant in the context of this research are discussed here, one looking at the perceptions of healthcare professionals regarding the unsuccessful cara.data sharing initiative and a more recent study with a broader focus, aimed at exploring the views of healthcare professionals about the secondary uses of patient data. Shortly after the Care.data initiative was cancelled, Ford and colleagues (2020) conducted aa study was conducted to understand the views of general practitioners about the scheme. Their work consisted in conducting content analysis of 162 media articles in which general practitioners expressed their views about the scheme. Findings of this work suggested that when Care.data was announced healthcare professionals supported it, recognising that this could be beneficial for research and could help improving health outcomes, nevertheless, as more information about the scheme became available this community of professionals expressed a number of concerns. For example, they felt that patients were not given sufficient details about the scheme (e.g. how data was going to be used and by whom), they also were concern because they felt unable to provide reassurance to patients that their confidentiality would be protected and that their data would not be used for profit. One of the key reasons that led healthcare professionals to express distrust in the scheme was that they were aware that in previous occasions the government mishandled sensitive data. They also expressed concern for the potential negative impact that this initiative could generate to the relation between healthcare professionals an patients, they felt that if they were not able to assure patients that their data would be adequately and safely handled, they could lost the trust of patients as some people would choose to not disclose important medical issues, this

naturally would impede GPs to provide adequate care which would damage certain groups, particularly those already in a vulnerable situation. General practitioners also expressed that the opt-out system proposed could be challenging for some vulnerable groups, nevertheless while they had this concern, they understood that an opt-in system would not have been the best option because that would generate bias and have a negative impact on the quality of research. The results of this work suggest that healthcare professionals are aware of the benefits of health data research and are willing to support it if adequate measures to handle data are in place and if the interests of patients are protected, however they are likely to express concern and withdraw their support to the reuse of patient data if they perceive that this could generate negative consequences for patients or damage certain vulnerable groups. A more recent study was conducted in 2019 to understand healthcare professionals' views concerning the secondary uses of healthcare data (Neves et al., 2019). This study explored knowledge on its purposes and the main concerns of healthcare professionals about data sharing processes in England. Healthcare professionals from different geographic locations were interviewed ("London, West Midlands, East of England, North East England, and Yorkshire and the Humber" (Neves et al., 2019, p. 1)). According to the findings of this study, participants demonstrated having a comprehensive understanding of how patient data is reused beyond the healthcare sector, including research purposes. They pointed out that healthcare professionals, in general, felt comfortable with frameworks governing data sharing and reuse, as well as with their application. According to the findings of this study, healthcare professionals tended to express concerns about the level of accuracy of data, the willingness of patients to share their medical records, and potential patient exposure or exploitation (Neves et al., 2019).

Among these concerns, the most significant for participants was towards the security of data and potential breaches. Healthcare professionals expect that external organisations reusing patient data will reassure them that they will ensure that the confidentiality of people will be protected at all times.

Another example of a study that has paid attention to non-patients or public stakeholders' attitudes is the work of Mouton and colleagues (2018). This research examined Swiss physicians, policymakers, and ethical committee members' views concerning the reuse of large patient datasets. This study is the first that has been conducted in Switzerland to explore this subject. According to their findings, healthcare professionals and administrators tend to be mainly concerned with data protection frameworks, security of data, mechanisms to ensure

confidentiality, and respecting the rights of patients. A key priority for interviewees was the use and implementation of legal frameworks for collecting and using data.

In contrast, they showed less interest in matters related to the agency of patients and shared decisions concerning the administration and uses of medical records. Interviewees of this study tended to view patients only as "data donors" (Mouton et al., 2018, p. 1) while overlooking the potential opportunities of significant engagement in decision-making. According to Mouton and colleagues (2018), the fact that health stakeholders seem to be mainly interested in scientific and legal matters concerning the production and reuse of patient data reveals a potential tension between the public interest and research priorities. For Mouton et al. (2018), rather than setting public interest against patient rights, it is necessary to develop and adopt approaches to strengthen both of them.

2.4.3 Information governance and the circulation of data

Other aspects concerning the circulation of data that have been explored in the literature, but received less attention include information governance, legal and ethical barriers for the circulation of data. For example, Vikström and colleagues (2019) conducted a study to examine barriers that obstruct the reuse of medical records' data for research purposes in Finland and Sweden. According to findings of their study, information governance procedures can act as barriers for the circulation of patient data for research purposes. They pointed out that in the Finnish context, the data request procedures were considered to be extensive and unstructured, and only generic instructions are provided by the data producer to support data requests. Due to the complicated nature of these processes, data reuses are often required to perform tasks that can be time consuming and therefore block or slow down the circulation of data.

Others that have paid attention to information governance aspects are van Velthoven and colleagues (2016) who examined information governance processes required for obtaining access to electronic health records' data across 16 countries. Their intention was not to identify approaches to overcome information governance barriers. Rather, their objective was only to gain an understanding on how information governance procedures operate across the nations studied. Findings from this study reported that obtaining permission for reusing data from electronic health records for research purposes was relatively simple in the majority of nations explored. However, there were some exceptions; for example in countries such as South Africa and India more barriers existed for obtaining permission, and in Austria requesting access was

not viable because in this country medical data were not meant to be reused beyond the healthcare sector (van Velthoven et al., 2016).

2.4.4 Conclusion

Key common findings and observations in the information studies and health literature with a techno-managerial orientation reviewed concerning data flows:

- Technical and data management are major barriers that can slow down or block data movement as they flow between people, organisations, or machines.
- Barriers that slow down or block the movement of data are commonly perceived as negative. Therefore it is considered that they should be overcome.
- Developing and implementation of best practices and strategies can help to overcome barriers to data flows.
- Data quality issues have been identified as one of the major barriers for the movement of data.

Key common findings and observations in the literature reviewed concerning health data flows:

- Datasets produced in the healthcare sector are often incomplete, fragmented and inaccurate.
- These limitations in datasets have motivated researchers in the field of health studies to develop and adopt approaches to remove barriers associated to quality issues.
- The following five dimensions are commonly assessed to determine the quality of data: completeness, correctness, concordance, plausibility and currency.
- Completeness seemed to be considered a key marker of quality and one of the most commonly assessed dimensions.
- Patients and public are willing to support the reuse of patient data for research purposes if this is for the public benefit and if confidentiality can be maintained.
- Whereas patients and the public seemed to be supportive of the reuse of patient data, one of the main concerns reported by healthcare professionals concerning the circulation of patient data beyond the healthcare sector is patients' unwillingness to share data.

Findings from this literature review reveal that the orientation of studies about health data flows in the information studies and health literature has largely been in relation to barriers for reuse and perceptions of reuse. Its main interest seems to revolve around two key aspects. First, to identify

and understand technical challenges that while participants do not point them out, can create frictions that slow down the movement of data or completely block data flows. And second, to understand the attitudes of people towards the flows of data beyond the healthcare sector to motivate research teams to stay within the boundaries of what is acceptable to most people.

2.5 Theoretical orientation: Critical Data Studies

This section discusses the emergence and key theoretical notions of the Critical Data Studies (CDS) field, which is the approach adopted in this study. The Critical Data Studies field is characterised by paying attention to ethical, political, power, and social issues related to data (Iliadis & Russo, 2016b). This field recognises the non-neutrality of data and data infrastructures (Gitelman & Jackson, 2013; Kitchin & Lauriault, 2014), and the materiality of data (Kitchin, 2014d). In a further section of this work (2.5.1), I will discuss how the aspects above relate to the Data Journeys approach (Bates et al., 2016).

Critical Data Studies arose in response to deepening processes of datafication and related claims about, e.g., the power of big data. The development of digital data infrastructures, as well as the expansion of big data, datafication, and dataveillance, poses a number of questions about the nature of data, the ways in which they are being created, organised, analysed and used, and what is the best way to make sense of them and to develop an understanding of the work they do (Kitchin & Lauriault, 2014). Critical Data Studies emerges as a field that strives to critically address these questions (Dalton & Thatcher, 2014; Kitchin & Lauriault, 2014).

CDS calls for reflection on the nature of data, taking into account the complex socio-technical system that they are part of. Such a system is integrated by a number of mutually constituted apparatuses and elements (Kitchin & Lauriault, 2014). Accordingly, Critical Data Studies scholars have stressed that to make sense of data, particularly 'big data', it is crucial always to situate data in time and space. It is also key to acknowledge that big data as a technology is not a neutral tool. Rather, it is always shaped, and it is shaped by a politicised scenario in both its construction and interpretation (Kitchin, 2014d). Data is always 'cooked' (Gitelman & Jackson, 2013, p. 2) through its creation and resulting interpretation.

Scholars from this field highlight that it is important to pay attention to power asymmetries between data creators, data captors and data analysts across time and in different contexts (Dalton et al., 2016). They have also suggested exploring issues around data ownership and data control (Kitchin, 2015). One part of the field seeks to study data assemblages to map the data landscape or understand how they are integrated and generate comprehensive descriptions

of them (Kitchin & Lauriault, 2014). Its intention is also to develop an understanding about how data are created, processed, and pervade and exert power in different contexts within the society (Dencik et al., 2016; Iliadis & Russo, 2016b). In this sense, CDS questions all forms of potentially depoliticised data science (Iliadis & Russo, 2016b). Contributors in the CDS field have placed the social at the core of data narratives, challenging technically deterministic assumptions that obscure the social character, consequences and impacts of data technology (Lee & Cook, 2020).

The field of CDS seeks to invite us to reflect about Big Data in terms of the public good and social settings and offer adequate resources to individuals for becoming better informed (Iliadis & Russo, 2016b). The ultimate goal is not only to map Big Data but to change it (Dalton et al., 2016). CDS seeks to provide tools to organise efforts aimed at developing ethical routes through a datafied world (Iliadis & Russo, 2016b). It is also interested in fostering the implementation of just practices concerning the ways in which people are made visible, represented and treated as a consequence of their production of digital data (Iliadis & Russo, 2016b; Taylor, 2017). To unpack the complex assemblages in which data are produced, shared and used, CDS proposes the application of social theory (Iliadis & Russo, 2016b; Kitchin, 2014d). Methods such as ethnographies, interviews, focus group and participant observation have been referred as suitable resources to conduct explorations of data assemblages (Kitchin, 2014a).

Scholars who position their research in Critical Data Studies, have paid attention to various aspects of the "life of data" (Bates et al., 2016; Burns, 2018; Currie et al., 2019; Mulder et al., 2016), aiming to illuminate some of these processes. For example, in the context of crisis mapping, Mulder and colleagues (2016) have drawn attention to the creation process of crowdsourced big data. They conducted a qualitative study to analyse the 'big data generating' process through crowdsourcing using open data platforms shortly after two humanitarian crises (the 2010 Haiti earthquake and the 2015 Nepal earthquake). This study pays attention to how data are created, and the series of transformations they undergo as they are processed and moved between different stakeholders who participate in the 'big data generating' process. In so doing, this work shed light on how and to what extent affected civilians can contribute to and access humanitarian crisis data (Mulder et al., 2016). A key contribution of this work is that it shows how the social processes of transformation that take place in combination with information transfer and translation in the process of generating humanitarian big data give rise to the exclusion of certain groups of people or communities (Mulder et al., 2016).

Another example of a study which has adopted a CDS approach to explore data in a related context is Burns's work (2018). Drawing from an ethnographic work conducted in New York City and Washington DC after Hurricane Sandy in 2012, this author paid attention to the processes that shape data production and representation. This work contributed to developing an understanding of the community-based and institutional politics that frame the types of data produced and the different ways in which those data are portrayed, in disaster contexts. This study shows that these politics emerge from the various contested data practices performed by digital and other types of organisations, formal disaster response agencies, and individuals. By paying attention to the politics and struggles around data practices, this work shed light on big data's social and political inequalities (Burns, 2018).

CDS researchers have addressed many such issues related to the political, ethical, and social issues related to data production and use. The focus of the next section will be CDS work on the circulation of data. Addressing the circulation of data has been an important focus in the big data context where the ways in which data move between different sites and organisations is dramatically changing.

2.5.1 Critical Data Studies research on data circulation and data journeys

A number of scholars studied the circulation of data before the CDS field flourished who might now be identified as key contributors to this domain. One of them is David Beer, a researcher within the Media and Culture studies field, who examined the different ways in which digital data moves across popular culture (2013b). His work paid attention to the infrastructures and artefacts that enable the generation of data through day to day interactions with popular culture. This author explored the circulation of data, the obstructions that data encounter, and the transformations they experience by examining the role that different elements such as archives and algorithms play in those flows. Beer's study (2013a) uncovers the systems and flows that shape what popular culture is, how it is structured, how it is spread, and how tastes and preferences are shaped or moulded. Key findings of his work suggest that emergent circulations of digital data play a key role in shaping what popular culture is and how it is experienced. This takes place as data in various forms accumulates and feeds back into the generation, distribution and consumption in multiple, and at times, unperceivable forms (Beer, 2013b). Beer's analysis was helpful to begin developing an understanding of how data are created, the ways in which they accumulate, how they are organised, and how they flow.

Other researchers who have addressed the circulation of data from what might now be considered a CDS approach are Bates and colleagues (2016). They conducted an empirical cross-disciplinary study focusing on the circulation of meteorological and climate data. For this study, the authors employed a methodology developed by them called *Data Journeys*. This methodology aimed to shed light on the socio-material composition of data and their movements as data navigate across diverse sites of practice. This approach centres on 'the life of data' (Bates et al., 2016, p. 2) as they travel across space and time through diverse cultures and sites where data practices occur. Data Journeys examines the flow of data across different sites of production, processing, dissemination and use. Bates et al.'s work (2016) uncovered how different sociocultural values and material elements coalesce through time to generate the socio-material conditions that shape data practices and as a consequence exert influence upon the form and utilisation of data objects and their circulation between data infrastructures. One of the journeys they studied was the recovery of historical climate data from archived ship log books to be used in global climate science databases. This empirical exploration (Bates et al., 2019) uncovered frictions in the composition of this data infrastructure and data movement through it. It also shed light on the adaptive and reflexive character of the practices and ideas encouraged by key stakeholders in the data assemblage to advance efforts to develop an infrastructure despite major challenges. Findings of this study showed how the desire and enthusiasm of key actors interact with the socio-material context in such a form that enabled them to defeat frictions restricting the movement of data and recover important data regardless of not having access to sustainable funding (Bates et al., 2019). They also identified that significant vulnerabilities emerged mainly because this initiative was developed in an adaptive way in relation to the neoliberal context, without engaging in a critical and profound reflection of these conditions (Bates et al., 2019). The key vulnerabilities identified by Bates et al. are the heavy dependence on voluntary work, a drive to address the climate crisis mainly using scientific and technological resources, and the absence of a critical view concerning the role of worldwide "neoliberal financial capitalism" (Bates et al., 2019, p. 15) in exacerbating environmental crises.

The Data Journeys approach (Bates et al., 2016) contributed to the development of an understanding about how different social realities are becoming increasingly interrelated and also how their interdependence is growing as they help to the formation and are influenced by nascent data practices (e.g. data generation, dissemination and use). It revealed the importance of sociocultural values formed in an historical way in framing data production practices and

showed how the material properties of data, such as their "volume, mutability, and durability affect how data move across different sites" (Bates et al., 2016, p. 10). By highlighting the wider dynamics of power shaping the evolution of different data practices, the Data Journeys approach sheds light on how data practices and flows are deeply politicised.

2.5.2 Empirical studies on data flows in critical data studies and related fields

Beyond key contributions on "data circulation" (Beer, 2013b) and Data Journeys (Bates et al., 2016), a number of researchers in CDS and related fields have conducted empirical studies aimed at exploring the circulation of data. In this section, these studies will be discussed in detail, ending with studies specifically adopting the Data Journeys approach used in this research. Whereas the studies here presented have adopted different approaches, they have some similarities. For example, they acknowledge the social construction of data and data flows as well as its social implications. The studies here agree that data are not neutral and that there are a number of factors that shape how they move and how they perform (Kitchin, 2014a). These studies centre their attention on issues such as the expectations of citizens in relation to how data is used, privacy implications of specific data practices, and the impact that the use of data can have for specific individuals or communities.

Some studies have explored data flows of data from its production through to use in policy. An example of a study of this nature is the work of Leite and Mutlu (2017). They conducted empirical research aimed at tracing the path of data through data production, dissemination and use in disaster response and migration policy to understand how data is collected, analysed, and contextualised in policy governance practices across the European Union. This empirical study (Leite & Mutlu, 2017) examined how data are converted into facts and the relationalities of this procedure in two areas under EU governance authority: provision of security concerning migration and mobility and disaster governance. This work sheds light on the role data play in transforming how policymaking works. It showed the ways in which phenomena are translated by data into social events that can be managed and controlled in a political way, enabling stakeholders to utilise data to define their relative roles within policymaking networks. A key contribution of Leite and Mutlu's (2017) work is that it highlights that data change how governing bodies work, thus demonstrating that data's social nature produces social effects.

An example of research that has paid attention to what citizens expect concerning the use of data is Beckwith and colleagues' (2019) work. They reflected on the relationship between cities and data, paying particular attention to the way that citizens expect data about their

communities to be managed. Their work proposed three conceptual frameworks for developing data governance policy for the people and discussed how, in their understanding, smart city data should flow. This study investigated a town located in the United States of America, which experienced considerable flooding problems in recent years. Over two years, Beckwith and colleagues (2019) worked with residents, government agencies involved with the community and a special interest group that was making efforts to shape funding and policy in the investigated town. These authors argue that smart cities have two options regarding governing and managing their data. They can opt for a model where information resources are open and available to all or understand and manage data as commons. Open data is commonly free to everyone, and there is no owner controlling how data moves. In contrast, data commons is expected to be treated as a resource held in common by a group.

Beckwith et al. (2019) suggest that smart cities can be the epicentre for creating new value for inhabitants of a town. However, they can also be places where data are produced and shared to benefit external stakeholders and at the same time, harm those within the city. For this reason, this study suggests that 'smart city data' should be treated as commons and may require some restrictions in terms of what and how data flows. According to Beckwith et al. (2019), cities should perform a social role of data stewardship of data both when the data remains within the boundaries of the city, and also in the cases where data travels towards external sites where different data practices are performed. From Beckwith and colleagues' perspective (2019), data openness may not be the best alternative for a community in all cases. This observation, similar to Bates' (2017) argument suggests that not all data frictions should be overcome. Instead, some of them are worth preserving.

Other scholars have centred their attention on the movement of social media data. Halford et al. (2018), for example, explored the ways in which Twitter or other types of social media data are framed for secondary research analysis. They drew attention to the socio-technical processes that shape the formation and flows of social media data. In this work, the potential challenges that can emerge when using social media data for research are discussed. Halford and colleagues (2018) proposed practical guidance to assist the design of a methodological approach for social media research. When using social media data for research these authors suggest it is important to report with transparency how data was harvested and to reflect about what constitutes data, what story they can present and what they do not show (Halford et al., 2018). They highlighted that social media data could contribute to understanding the social world, but researchers are encouraged to be careful not to make claims about their social significance. This idea echoes one

of the key observations made by contributors from the CDS field: data are always partial representations of the world (boyd & Crawford, 2012). With this work, Halford and colleagues contribute to extending the theoretical approach to data infrastructures into a useful approach to social media data methodology (Halford et al., 2018, p. 10).

The circulation of data in educational contexts has also received significant attention. Hartong and Foerschler (2019) explored digital data infrastructures, circulation, and practices on school monitoring in the public education field. This work paid special attention to what and how data is used, and the consequences of its use. This empirical study examined the ways in which datafication and digitalisation of the governance of schools have displayed within and across school settings and systems in state educational bodies in the United States and Germany (Hartong & Foerschler, 2019). Findings of this study demonstrate that in a similar way as in other contexts, data practices in the education sector always involve political implications, mainly when applied to accountability systems (Hartong & Foerschler, 2019). According to these authors, monitoring infrastructures created limited representations of the student, the teacher or the school. These representations lead to creating new categories and inaccurate representations of the analysed object that managers employ to communicate on "behalf of 'students at risk' or schools 'in need'"(Hartong & Foerschler, 2019, p. 10). In addition, such representations highlight some things while suppressing others, which ultimately may lead to the transformation of the key characteristics of what is being examined. This study suggests that datafication in state education agencies is mediated through practices and material elements (infrastructures) that operate in combination to execute calculation, assessment, and data work. These authors emphasised that "data practices always have political implications" (Hartong & Foerschler, 2019, p. 10) and also that data infrastructures are "always expressions of knowledge/power" that influence the types of questions that can be posed, in which way they are framed, the ways in which are addressed, the ways in which the responses unfold and who is able to ask them (Hartong & Foerschler, 2019, p. 10).

Another example in the education context is the contribution of Selwyn and colleagues (2015). They conducted an empirical study in secondary school settings to illuminate the ways in which administrators, managers and senior leaders were experiencing digital data and data work. This empirical study paid special attention to what data was being accumulated in schools, how these data was being used, and the consequences of these data's uses. Their study focused on the use of digital data among the staff of public secondary school settings in Melbourne, Australia. Findings show that most individuals within the school communities were engaged

in unequal ways with data, which seemed to reproduce broader administrative and managerial hierarchical structures within the schools. Data in the school settings examined was being generated and processed so that the historical character of the secondary schools or the people that were part of them was being erased (Selwyn et al., 2015). That is to say, that data was only being used to create decontextualised stories about these schools. This study unveils that digital data can be utilised to transform power into something even less perceivable and implicit. Selwyn and colleagues stressed the need for raising awareness among stakeholders within secondary schools about how digital data are used "'on them' and 'for them'" (2015, p. 780). This work draws attention to the social construction of data, acknowledging a key theoretical assumption from the CDS field, which is that these objects cannot be taken as neutral facts as they represent choices and interpretations (Kitchin, 2014a).

The flows of young children's data have also been explored adopting a CDS approach. For example, Roberts-Holmes and Bradbury (2016) examined the datafication of the young children's education sector in England, intending to make visible the circulation data. They examined the way in which comparative data-based accountability progressively managed the pedagogies and professional practice of early years (e.g. nurseries and primary schools) teachers. Their findings unveil the ways in which datasets that are meant to be used for comparison are increasingly being used to manage local educational practice (Roberts-Holmes & Bradbury, 2016). This study also shows how the use of these datasets provoked a shift in the locally contextualised professionalism towards generating objective numerical data that could be compared within a centralised system managed at a national level. According to Roberts-Holmes and colleagues, these dynamics may result in young children being seen as the source of the statistical material that is to be extracted and exploited to achieve the school's goals. Smith and Shade (2018) are others that have shown interest in exploring data flows of young children. Their work paid attention to the privacy issues concerning the use of children's data in digital playgrounds. These authors traced children's data flows by conducting a document analysis of privacy policies and other key documents publicly available of YouTube Kids and the Fisher-Price Smart Toy. This study analysed the commercialisation and dataveillance practices that transgress children's and parents' privacy rights as children's data travels through surveillant playgrounds (Smith & Shade, 2018). Findings show that companies expect that parents or legal guardians assume the entire responsibility of controlling and supervising digital playground profiles accessed and utilised by children aged 13 or younger. This study also uncovered how children's data is analysed to offer them personalised content and to understand

how the "child is learning" (Smith & Shade, 2018, p. 9). Furthermore, this research reveals that the ambiguous language used in playgrounds' privacy policies may confuse some parents. This work highlights that the collection and use of children's data in platforms and apps raise privacy concerns and worrying ethical tensions, which include device security.

2.5.2.1 Empirical studies influenced by the data journeys approach

Of the various empirical CDS studies that have examined data flows in different contexts; some have adopted the same Data Journeys (Bates et al., 2016) approach that this study adopts. Swist and colleagues' (2019) focus was on children's data. Their work explored the flows and frictions of data from children in the digital era and the ways in which datafication works as a form of understanding and governing children. Swist et al. (2019) paid attention to the ways in which 'the child' is created through the circulation of children's data by swiftly multiplying digital paths, intersecting roads and units of storage, which are ingrained in propositions and ambitions focused in the adult, interfering with the agency of children and control over data about them. These researchers employed the data journeys methodology to explore the different infrastructures and interests of three types of organisations: corporations, governments and universities, with the intention of uncovering the distinct paths by which children's data are unevenly gathered and used. By documenting the journeys of children's data, Swist et al. (2019) were able to uncover the ways in which data infrastructures and interests can ignore, obfuscate, or take advantage in an unfair way of children's personal experiences and agency, leading to the creation of the child as a 'data point' in a datafied economy focused on adults. They showed how pre-established top-down paths minimise difference and children's lived experiences with almost no constraints, which further challenges data-driven approaches' reliability and effectiveness (Swist et al., 2019). Their findings also show that the journeys of children's data trouble the supposed smoothness of data movement and emphasise different frictions generated by the "datafication of the child" (Swist et al., 2019, p. 73). Drawing from the insights gained through this exploration, Swist and colleagues proposed redirection of academic, companies, and government approaches to data and the child towards networked capability and child rights frameworks. A child rights approach attempts to give a route to traverse the intricate landscape of data governance, whereas the networked capability approach intends to be helpful to set direction to explorations that consider children's differences and values.

Another author who drew from Bates et al.'s ideas to explore data flows was Ville Aula (2019). To date, this is the only study on health data using the Data Journeys/CDS approach. His work

examines the national reform of health and biomedical data infrastructures and associated legislation (referred to by the author as Secondary Health Data Initiative) that took place from 2014 to 2019 in Finland. Aula's work explored the role that data governing state bodies play in reforms that attempt to make data flows smoother and minimise data frictions, and how data frictions emerge at institutional and regulatory levels. Aula's empirical exploration draws from previous work that has paid attention to the circulation of data (Bates et al., 2016; Edwards et al., 2011; Leonelli, 2016). He considers Leonelli's (2016) data journeys work situated in the STS domain, Bates et al.'s (2016) data journeys approach, which can be situated in the CDS field, and the data frictions notion developed by Edwards and further developed by Bates. Empirical findings of this work show that since the Finnish Secondary Health Data Initiative's objective was to enable a more smooth circulation of data between governors and collaborators, the project's practical side was centred on minimising the frictions between data infrastructures and institutional jurisdictions (Aula, 2019). It also demonstrated that Finnish institutions governing health data performed a key role in the existing frictions in the movement of data. However, frictions also emerged during legal negotiations concerning expectations regarding the desirable state of the health data infrastructure and their governing (Aula, 2019). The findings also show that whereas data practices and organisational elements are linked, it is less challenging to change data practices than regulatory and governance practices as well as power dynamics across institutions. Aula's work (2019) highlighted that government institutions could play a key role as guardians of the interest of the public in the context of big and open data motivated reforms.

While some researchers have adopted the Data Journeys approach as described, others have adapted to their own requirements. For example, Merricks (2017), drawing from the methodology developed by Bates et al. (2016), proposes to follow data from their initial creation through their reuse in different settings. Merricks however, reconceptualises Bates et al.'s approach (2016) and uses a different metaphor, proposing to follow "data threads" rather than "data journeys", in an attempt to highlight that data infrastructures and geography are intertwined and inherently related (Merricks, 2017). Using the data threads approach, Merricks (2017) conducted an empirical study focused on infant mortality statistics data. In this work, he followed infant mortality data generated in Toronto from their initial creation through their different uses, drawing from interviews conducted with public officials in Toronto, Canada, and primary and secondary sources of health data also produced in this country. Findings of this work show that the adoption of international standards used to record mortality data

excludes or nullifies the feedback that local domain-area experts develop on an ongoing basis (Merricks, 2017). It also reveals that the decisions of those in charge of determining what data is recorded and shared for international comparison have led to the lack of representation of some areas. This means that infant mortality statistics are, on the one hand, internationally compliant inputs, and on the other hand, inaccurate representations of infant mortality (Merricks, 2017). This study also uncovers how as infant mortality data move off from their creation point, they become increasingly abstracted and estranged from the initial event they represent. Data threads is presented as a useful approach to explore issues of provenance, licensing and veracity (Merricks, 2017).

2.5.2.2 Conclusion

Key common findings and observations present in the empirical work on the circulation of data include:

- Data are socially constructed, therefore, cannot be understood as neutral facts.
- The social nature of data produces social effects.
- Datasets are always partial, incomplete representations of the world.
- Power dynamics play a significant role in shaping data flows.
- Data infrastructures are depictions of knowledge/power that influence the types of questions that can be posed, the ways in which they are addressed, how these answers unfold and who can ask them.
- Data practices always have political implications.
- The circulation of certain types of data (e.g. health and children data) raises privacy and ethical concerns.
- Not all barriers for data flows should be overcome. Some restrictions for the data to move are worth fostering.

Despite this wide range of CDS and related studies on data circulation, little attention in the CDS field has been given to study health data flows and reuse in the context of the UK's NHS. Within this field Aula's (2019) work is one of the few studies that have paid attention to data flows in the healthcare sector. However, he centres his attention in the infrastructural and legal reform of secondary uses of data in Finland, paying specific attention to the role of data governing state bodies shaping data flows.

2.6 Data Journeys: a critical data studies approach to illuminating the socio-material constitution of data flows

As addressed in previous sections of this chapter, researchers have adopted multiple approaches to the study of the circulation of data. Such approaches have been shaped by a number of factors, such as disciplinary and philosophical orientations. For example, some studies appeared to be more descriptive whereas others tended to be more critical. In terms of epistemological orientations, some of them were interpretative whereas others adopted more critical approaches. Another key difference is that some have sociological whereas others have more instrumental ends.

This study adopted the Data Journeys (Bates et al., 2016) approach, which as mentioned before, is considered to be part of the Critical Data Studies field. Bates et al.'s (2016) data journeys approach was developed as a methodology to examine the socio-material constitution of data and flows as data moves across diverse cultures and sites of data practice, from their generation point through to reuse in different sectors. Data Journeys places particular attention on the multiple social worlds which are connected to each other, partly, by the movement of data through and across distinct sites of practice. It has the objective of uncovering the specific ways in which different sociocultural values and material elements coalesce through time to generate the socio-material conditions that shape data practices (e.g., data production, processing, and distribution) and as a consequence exert influence upon the use of data and their movement between infrastructures (Bates et al., 2016). The approach requires that the researcher follow data as they move through interlinked organisations and projects (Bates et al., 2016). It pays attention to how and why data moves from one point to another and issues related to, e.g., obstructed movement and lack of movement. While data journeys can be used to examine the socio-material constitution of both data objects and flows, in this study only the socio-material constitution of data flows will be examined.

2.6.1 Key theoretical assumptions of the Data Journeys approach

In this section I will describe the key theoretical assumptions of the Data Journeys (DJ) approach and define key concepts that the DJ approach draws on. Details about the methodology will be addressed in the methods chapter.

2.6.1.1 The materiality and non-neutrality of digital data

The data journeys approach aims to illuminate some of the ways in which data objects and flows have a socio-material constitution. This approach was developed in agreement with considerations that point out that data are not neutral but instead are socially constituted, material objects. Common assumptions that suggest that data are neutral, objective, and pre-analytic in nature or that define them as only the 'building blocks' or the 'raw material' to create information and knowledge have been labelled as limited (Lauriault, 2017). These ideas have been contested by a number of scholars who understand data in a much more complex way (Borgman, 2015b; Gitelman & Jackson, 2013; Kitchin, 2014d). Scholars who reject the idea that data are independent objects highlight that data do not emerge from nowhere and their generation is not spontaneous (Kitchin & Lauriault, 2014). Rather, their production is a result of human action (Edwards, 2013; Iyamu & Mgudlwa, 2018). Therefore, data cannot be understood as neutral and objective because they are always infused with the assumptions and politics of their creators and calculators (Merricks, 2017); they are not simple evidence of phenomena, they are phenomena in and of themselves (Wilson, 2015). In Gitelman and Jackson's (2013, p. 2) words, they are "'always already 'cooked' and never entirely 'raw'." In this sense, it can be said that the world is not just reflected in data (Ribes & Jackson, 2013) but data perform active work in the world (Kitchin & Lauriault, 2014, p. 11).

In the data journeys approach, socially constituted digital data are recognised as material objects. This notion that data are material objects is relatively common in the literature of CDS and related fields (Lupton, 2015). For Bates et al. (2016) the materiality of data can be perceived in various ways, for example, contained in the "magnetic atoms of a hard drive and when they send wirelessly as electromagnetic signals" (Bates et al., 2016, p. 3). Data materiality can also be identified through considering that data are the fruit of a series of practices through which cultural values come into being in the shape data take. The Data Journeys approach also considers that data have material consequences.

Bates et al. (2016) employ a set of definitions to refer to different aspects of the materiality of data and their movements; these terms include:

- **Physical infrastructure:** to make reference to "material objects (artefacts) such as computer and instruments" (2016, p. 3).
- **Physicality of data:** to make reference to "the atomic and electromagnetic form of data" (2016, p. 3).

- **Crystallisation:** when addressing the ways in which "sociocultural values take substance in the materiality of data and their infrastructures" (2016, p. 3).
- **Material:** this term is used in two different ways: 1) to refer to "the material conditions of production" (2016, p. 3), and 2) to refer to "the material properties of data" (2016, p. 3).

2.6.1.2 *Researchers must consider the nature and aim of a study to decide which of these aspects of the materiality of digital data and their movements are more relevant to examine. In the case of this research the emphasis is on the material aspects of data movement, rather than data as objects. The material aspect of data flows that is explored is how material factors (infrastructure and material conditions of production i. e. funding, investment, policy) shape the circulation of health data.Issues concerning the physicality or material properties of data are not explored. I specifically pay attention to how these factors interact and work together with sociocultural factors to drive the movement or produce sources of friction for patient data to move towards universities to be reused for research purposes.****Defining sociocultural factors***

Whereas Bates et al. (2016) emphasised the importance of sociocultural factors in the socio-material constitution of data, they do not provide a definition of what specifically they understand sociocultural values are. However, from what they explain in their work exploring the socio-material factors that shape the formation of an infrastructure for weather data (Bates et al., 2019), it can be said that they understand sociocultural values as : 1) motivations for doing something or engaging in a task or activity, 2) cultural features shared by a group of people or a group of professionals, and 3) desires, affects or wishes.

Bates (2018) goes on to expand on what is meant by 'sociocultural' in relation to data practices in a later paper. She argues that the 'data culture' of a site of practice can be understood as the "different cultural norms, value systems and beliefs that inform, frame and justify people's data practices" (Bates, 2018). The data cultures of sites of practice can be illuminated if philosophical beliefs, sociocultural values and assumptions, and the perceived value of data practices are critically examined (Bates, 2018).

According to Kapofu (2021) in order to make sense of the culture of a group different components of the sociocultural can be explored. This includes for example, 1) what is observed by outsiders, including components such as clothing, special distribution and

placement, technologies, and archived materials, 2) what is produced by insiders and is integrated by artefacts that are only produced and reproduced by the group, or in other words, the components that hold the group together, for instance language patterns and social norms; 3) what is professed to outsiders, that is to say what a group embrace and support and wants outsiders to know about them, this includes what and how a group do what they do, and; 4) what is meant by insiders. Exploring this dimension is useful to understand why the group exists.

In this research when I talk about sociocultural factors, I am specifically referring to the following : 1) motivations for doing something or engaging in a task or activity, 2) social norms 3) beliefs and expectations, and 3) desires, affects or wishes, all of them shared by a group of people or a group of professionals. Examples of sociocultural norms and values of a group include their motivations for engaging in certain data practices, patterns of decision-making, internal dynamics, perceptions of their own culture, and ways in which they relate and engage with external cultures.

2.6.1.3 Data infrastructures

Data journeys propose to follow data through various interlinked organisations and projects within and across data infrastructures. In their approach Bates et al. (2016) acknowledge the social and relational nature of data infrastructures. Data infrastructures are understood as the organisational, physical and digital means for preserving, disseminating, and making use of data across networked technologies that at the same time are ingrained within broader institutional scenarios (Kitchin, 2014b, 2014d). According to Star (2010), data infrastructures are constituted by social and cognitive structures. On the one hand, social structures are integrated by the social and economic relations of individuals, state agencies, private organisations, and other actors that produce, analyse, disseminate and use data. On the other hand, the cognitive structure refers to the way in which data is structured by people (this includes the boundaries of inquiry, presumptions in relation to social reality, schemes of classification, measurement methods, and official rules for interpreting and presenting data).

The complex constitution of data infrastructures reveals that in the same way as data, they lack neutrality, and actually they are material manifestations of some broader social and political dynamics (Harvey & Knox, 2012). Data infrastructures are not just objective technical means used to gather and share data (Kitchin, 2014d), rather they are framed by a number of histories, ideologies, and philosophies, which frequently remain veiled. Also it is important to

acknowledge that they are not static units (Furlong, 2011). The ways in which data are processed and analysed within data infrastructures is transformed over time, and is affected by a number factors such as organisational change, laws in relation to data handling and data protection, and the development of new technologies (Kitchin, 2014a).

Infrastructures are one component of a data assemblage; however there are many other elements as well. As defined by Kitchin, a data assemblage is integrated by a set of co-functioning loosely connected material and social elements that shape data collection, production and sharing practices (Kitchin, 2014d). The elements of a data assemblage are on the one hand, the components of a data infrastructure, and on the other hand, the institutions and the environment within which data are situated (Kitchin, 2014d; Lauriault, 2017).

Data assemblages are deeply intertwined economic, political, social, and technological apparatuses and components (Iliadis & Russo, 2016b; Kitchin & Lauriault, 2014). According to Kitchin, the apparatuses that integrate a data assemblage are: finance, forms of knowledge, governmentalities and legalities, marketplace, materialities and infrastructures, organisations and institutions, practices, political economy, places, subjectivities and communities, and systems of thought (Kitchin, 2014a, p. 25).

Bates (2017) builds on the notion of a "data assemblage" in her conceptualisation of data friction in data journeys to create an analytical framework that centres attention on interrelations of infrastructure, sociocultural and regulatory factors, all of which are underpinned by political and economic dynamics. This framework offers a categorisation of three overarching elements that generate "friction" in the movement of data. These categories are (Bates, 2017, p. 416): 1) data sharing infrastructures and management, which consider for example, "technical infrastructures, data management practices, organisations, and materialities"; 2) sociocultural factors, such as "systems of thought, forms of knowledge, subjectivities, communities, and institutions"; and 3) regulatory framework, that includes for example "legalities, policy, and standards".

2.6.1.4 Adoption of the term 'journey'

Bates et al. (2016) propose to use the term 'journey', which highlights that data do not always move smoothly and free of their origins across different sites of practice. The term 'journey' is useful to locate data in physical space. In addition, the term 'journey' does a better job than the term 'flow', which tends to suggest a smooth movement without constraints. 'Journey' better

symbolises the ‘disjointed breaks, pauses, start points, end points’ (Bates et al., 2016, p. 4) and “friction” (Edwards, 2010) that take place as data circulate between diverse data infrastructures.

Bates (2017) further elaborates on the “data friction” concept in relation to the data journeys approach to illuminate the sociocultural struggles and politics around data flows in the cases of publicly funded research data and online communications data. However, she contests the idea that since data friction is an obstacle for scientific advancement, data frictions must always be overcome. She highlights on the contrary that friction is not always something that needs to be avoided or overcome, as in some cases it is worth it to permit and even foster it, for example in the case of privacy or ethical concerns.

Overall, data journeys is a helpful approach to uncover: 1) how data are processed and reused within diverse sites of practice; 2) how sociocultural values and material factors assemble to shape and justify practices of data reuse; and 3) how in combination these sociocultural values and material factors contribute to the generation of the socio-material conditions that shape data flows.

2.6.2 Justification for adopting the Data Journeys approach in this study

The considerations previously addressed in this section highlight that data are both social and material in the sense that they are situated in a context, and that they have a form (as numbers, symbols, bits, etc.) (Gitelman & Jackson, 2013) and are always framed and used in a context with the objective of achieving determined goals (e.g., to discipline, to empower, to regulate, to control, to produce profit, etc.) (Kitchin, 2015). The insights that emerge from data are not absolute truths; also they are not self-evident, rather they are constructed and interpreted (Tanweer et al., 2016). Since it is impossible for data to exist disconnected of the ideas, practices and contexts that intervene in the generation, processing and analysis of them (Bowker, 2000; Gitelman & Jackson, 2013), their value and meaning cannot be examined if we try to understand them in isolation (Borgman, 2015b).

The above theoretical considerations will guide this research and provide the means to explore the sociocultural dynamics shaping the journey of selected types of personal data produced within the UK healthcare sector.

As stated above, the CDS approach pays attention to power, political, ethical and social issues related to data and intends to develop an understanding of how data are produced and used and permeate and exert power in different contexts within society. In this sense, the adoption of a

Critical Data Studies approach will be useful to address the social and ethical concerns related to the practices of data processing and use of data generated in the UK healthcare sector.

The data journeys approach was selected among other approaches as it was considered helpful to achieve the aim of this research. The main reasons for selecting this approach are explained below:

1) In agreement with ideas proposed by key thinkers of the flourishing CDS field, the Data Journeys approach acknowledges the socio-material constitution of data and data flows and recognises that data and data flows have consequences. Since this understanding is a central component of the data journeys approach, it draws attention to interrelated material and sociocultural factors that play a role in shaping data and data movements.
2) This approach considers that data may encounter barriers as they move across sites of data practice and infrastructures. It is thus useful to identify 'breaks, pauses, start points, end points' (Bates et al., 2016, p. 4) and 'frictions' (Edwards et al., 2011) of data. This approach, however, does not prompt the researcher to identify barriers or constraints in the movement of data with the objective of always overcoming them. Rather, it invites the researcher to reflect on the politics and implications of those frictions, as it suggests that some frictions are worth preserving rather than overcoming.
3) This approach is based on following data as they move across different sites of data practice, suggesting the researcher should navigate the sites where data are processed or used, paying attention to the surroundings of data and absorbing the culture. This is especially useful to capture sociocultural factors that shape the movement of data.

2.7 The use of the concept of data valences to explore the circulation of data

In addition to adopting the data journeys methodology, the concept of data valences was used in this research to study the circulation of patient data. This section explains the concept of data valences, the reason why it was adopted and how it was used in the context of this research.

2.7.1 The concept of data valences

The conceptualisation of data valences, proposed by Fiore-Gartland and Neff (2015), aims to "describe the differences in expectations and assumptions that people have for data across different social settings" (2015, p. 1466). Data valences encompass a range of expectations and values for data that people convey through their discourses and practices. This conceptualisation can be particularly helpful to examine how data is valued and interpreted,

and what is expected from them across different contexts as well as to develop an understanding on how data work and perform in different social settings (Fiore-Gartland & Neff, 2015). The data valance concept enables researchers to study the way in which data are "rhetorically evoked" (Fiore-Gartland & Neff, 2015, p. 1470) and the ways in which discourses, dialogues, practices, and contexts of data deviate and reproduce even when it is considered that data interpretations are fairly consistent. Drawing from findings of their empirical work, Fiore-Gartland and Neff defined six data valences: "self-evidence, actionability, connection, transparency, truthiness, and discovery" (2015, p. 1466), recongising that this was not an exhaustive list and inviting other researchers to conduct empirical exploration in other contexts to define additional data valences. Building on Fiore-Gartland and Neff's work on data valences and the analysis of data collected I developed and proposed and additional valence, distinct to those proposed by them (2015) – the vanguard valence.

2.7.2 The use of data valences in this research

In this thesis the notion of data valences was specifically used as a lens to guide the analysis about how the expectations that university-based researchers have for patient data produced in the healthcare sector, influence the circulation of these data from the healthcare sector towards universities to be reused for research purposes. As follows the six data valences proposed by Fiore-Gartland and Neff are explained followed by an additional valence that emerged from the findings of this research, which is the vanguard valence.

The self-evidence valence refers to the notion that data are 'premade' resources that require neither work nor interpretation. The truthiness valence illustrates "how people expect data to comprise a single, direct, objective representation of a measured reality" (Fiore-Gartland & Neff, 2015, p. 1476). The discovery valence depicts how people expect data to be the genesis of discovering phenomena, issues, relationships, or states that otherwise would remain unknown (Fiore-Gartland & Neff, 2015). The actionability valence 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, p. 1474). The transparency valence is evoked when people talk about the benefits of "making data accessible, open, shareable, or comparable across cases and contexts" (Fiore-Gartland & Neff, 2015, p. 1475). The connection valence shows the expectation that data can be an excuse for engaging in meaningful conversations or an opportunity for making a connection with people (Fiore-Gartland & Neff, 2015). The additional vanguard valence, that I propose drawing from the findings of this research

illustrates how university-based researchers and some key stakeholders in the healthcare landscape perceive conducting research with data as the most innovative way of exploring health issues. The data valences identified in this work are discussed in the empirical chapers of this thesis.

Expected contribution of this research

As findings of the literature reveal, little research has paid attention to social, cultural, political and ethical issues related to the ways in which power dynamics play out in digital data production, reuse and movement. Therefore, this research aims to contribute to the body of literature on health data flows by answering the following research questions:

Main research question

- **RQ1** How do sociocultural values interact with existing and emergent material conditions to generate drivers and frictions that work together to produce the material forms of data journeys in the context of reuse of NHS patients' personal data for research purposes?

Sub-Research Questions

In order to answer the main research question, the following sub-questions are posed:

- **RQ2** What material forms do the 'journeys' of NHS patients' personal data take between different sites of practice, from their initial generation though to reuse for research purposes in different contexts?
- **RQ3** In what ways do sociocultural values interact with existing and emergent material conditions to act as drivers to move data between different sites?
- **RQ4** In what ways do sociocultural values interact with existing and emergent material conditions to act as frictions on efforts to move data between different sites?

Chapter 3. Methodology

3.1 Introduction

Methodology is not only a series of structured procedures or steps that can be utilised randomly and without any consideration to any empirical problem (Alasuutari et al., 2008). Methodology integrates an entire range of strategies and courses of action that include:

- producing an illustration of the empirical world;
- making enquiries about that world and transforming these into problems that are possible to investigate;
- figuring out the best ways of doing so — that includes decisions about methods and the data to be searched, the development and utilisation of concept, and the process of interpreting findings (Blumer, 1969).

Therefore, methods which are a "cipher for underlying philosophical ideas" (Bryman, 2008, p. 8) constitute only a small element within the complexity of research methodology (Alasuutari et al., 2008).

This chapter discusses the methodology used in this study and the rationale that supports the selection of procedures and techniques. This chapter is divided into five subsections. First, the research philosophy adopted will be explained, including philosophical assumptions (ontological considerations and epistemology). In section 3.3, the research approach of this study is presented. Section 3.4 describes the Data Journeys approach to research design and explains how it was used in this study. Section 3.5 describes the data collection of this study, which consisted of two phases. Afterwards, in section 3.6, the data analysis is explained in detail. Section 3.7 addresses research quality aspects of this study. Finally, ethical considerations and issues related to this study are discussed.

3.2 Research Philosophy

Research philosophy is a concept that is related to the construction of knowledge and its nature. The research philosophy adopted by a researcher incorporates significant assumptions and beliefs about the way in which they perceive and interpret the world (Saunders et al., 2015). These include ontological assumptions, which are concerned with the nature of reality or the nature of what exists, and epistemological assumptions, which are concerned with the way in which knowledge can be generated (Saunders et al., 2015).

It is crucial to understand philosophy in research, because it influences how a researcher interprets a problem, formulates a research question to investigate, and selects the research strategy to answer the question (Huff, 2009; Saunders et al., 2015). A researcher will always bring particular beliefs and philosophical assumptions to their research (Creswell, 2013). The philosophy adopted by a researcher will be influenced by practical considerations; nevertheless it is probable that the major influence will be the particular view that the researcher has about the relationship between knowledge and the way in which it is generated (Saunders et al., 2015). In the following sections, the underlying philosophical assumptions of this research are explained, as well as how these shaped and informed the adoption of research approaches, strategies, design and method.

3.2.1 Research philosophical assumptions

The philosophical assumptions that support research are highly important as they will determine its credibility (Farquhar, 2012). Ontology and epistemology have been extensively recognised as the main philosophical assumptions in social sciences (Bryman, 2012). The focus of ontology is on understanding what is, whereas epistemology seeks to grasp what it means to know (Gray, 2014).

3.2.2 Ontological considerations

Ontology refers to the study of the nature of reality and its characteristics (Creswell, 2013). Ontological assumptions can be made about what type of things do or can exist, the circumstances of their existence, and how they are related (Lewis-Beck et al., 2004). A number of authors within the Critical Data Studies field, including Bates et al. (2016) in their Data Journeys approach, have adopted a socio-material understanding of reality. In harmony with this approach, this research aims to gain understanding of how sociocultural values and norms interact with existing and emergent material conditions to shape "data journeys" (Bates et al., 2016) of patient data in the UK health sector, with a specific focus on reuse of patient data for research purposes. This study therefore recognises the importance of materiality (Iliadis & Russo, 2016b). Thus, in this research data are understood as objects. As explained in section 2.5.1.1, the material nature of data becomes evident in different ways, for example in the atoms of data's means of storage and in the electromagnetic signals through which they travel. The materiality of data is also perceivable in that they are the fruit of different practices through which cultural values are reflected in the shape data take. Different aspects concerning the

materiality of data can be explored; this work pays particular attention to how material factors (e.g. infrastructure and material conditions of production) influence the circulation of data.

This socio-materialist ontology differs from an idealist approach, which is characterised by only paying attention to the shaping power of things such as beliefs, discourses and ideas. It recognises the importance of the material world as well as these discursive things, and is interested in how different socio-material factors interact to shape the particular aspects of reality – in this study, how data objects circulate and how they are reused.

3.2.3 Epistemology

Epistemology is one of the central areas of philosophy (Given, 2008), which focuses on answering how we know that we know something (Miller & Brewer, 2003). There are three fundamental questions in epistemology: "what constitutes knowledge?", "what is the relationship between the knower and the known?" and "in which way knowledge claims are justified?" (Coghlan & Brydon-Miller, 2014, p. 303). Whereas there is a broad range of views around the focus of epistemology, "what counts as knowledge?" is considered the principal question in epistemology (Mathison, 2005).

Epistemology enables the researcher to have a philosophical background to determine what types of knowledge are valid and acceptable in a field of study (Gray, 2014). It can be helpful to clarify matters of research design, including the design of research tools, the type of evidence that is being collected, from where, and in which way it is going to be interpreted. Moreover, an epistemological perspective will allow the researcher to recognise which designs can work and which not for a particular set of goals (Gray, 2014).

In this research, a Critical Data Studies approach was adopted. This field has made a significant contribution to discussions about the epistemological implications of Big Data (Stevens et al., 2018). The following considerations are central to the epistemological critique of CDS regarding positivist views about big data systems:

a) Data are not neutral and objective representations of the world. This field has contributed to advance the argument that data are not neutral, objective, and pre-analytic representations of the world (Gitelman & Jackson, 2013). Rather, data are generated in complex contexts that are influenced by systems of thought, norms and regulations, organisational processes, technical instruments, economic forces, etc. (Kitchin, 2014d; Kitchin & Lauriault, 2014).

b) Outputs of big data analytics are not neutral representations of the world. In a similar way to data, the analysis of data and interpretation of patterns and relationships in data lack in neutrality as these are human actions infused with biases, assumptions and contexts (Kitchin, 2014d).

c) Data are always limited. Critics in the CDS field have drawn the attention to the fact that despite the large volume of datasets, big data can only offer limited representations of the world (boyd & Crawford, 2012). They also highlighted that it was crucial to recognise that regardless of how large in volume they are, they will always be incomplete in relation to any phenomenon the observer intends to explore.

d) Big data as technology for knowledge production is theoretically driven. Big data systems are not free of theory and philosophy; instead of this, they were designed with a specific objective and driven by an agenda, and their design is underpinned by scientific arguments and established theories (Kitchin, 2014d).

Building on these epistemological commitments, Critical Data Studies approaches are therefore non-positivist. Epistemological approaches that stand in opposition to positivism include interpretivism and critical. Interpretivism considers that reality is socially constructed and infused with multiple meanings (Mathison, 2005). This approach therefore explores different perspectives and pays attention to how people interpret the social world and social phenomena rather than simply observing facts (Matthews & Ross, 2010). In contrast to positivism, which seeks to develop abstract generalisations, interpretivism seeks to develop theories with the objective to provide an understanding of direct lived experiences (Mathison, 2005).

This study is inclined towards the interpretivist side of the CDS field, which means that it pays attention to "the meanings that humans connect to their actions" (O'Reilly, 2009, p. 119) when examining the reality. Furthermore, while it is also interested in power relations, it does not aim to grasp and examine them in such a deep way as more critical studies would (e.g. those using critical frameworks such as Marxism, feminism, or social justice). Nonetheless, it accepts key assumptions from more critical epistemological positions, which strive to challenge worldviews as well as the power structures that create them (Bronner, 2011). The epistemological position adopted in this research considers that the society is shaped by power structures (e.g. culture and politics) (Howell, 2013). It also considers that power relations influence and mediate ideas, and also that social relations play a key role in mediating both ideas and objects (Howell, 2013).

3.3 Research approach

The epistemological position of researchers will influence the way in which they work with theory (Matthews & Ross, 2010). Theories, understood as sets of ideas that intend to explain the social world, are significant to research methodology and the characteristics of the data a researcher gathers.

3.3.1 Inductive approach

In this research project, the aim is to gain understanding of the ways in which sociocultural values and norms interact with existing and emergent material conditions to shape "data journeys" (Bates et al., 2016) of patient data in the UK health sector, with a specific focus on reuse of patient data for research purposes. This research is exploratory, therefore the most suitable approach to address the research problem it proposes is inductive, rather than deductive.

In this approach relationships or theories are developed only after the researcher has collected and analysed data (Gray, 2014). The inductive process does not ignore pre-existing theories or ideas when addressing a research question; rather it does not intend to corroborate or falsify theory as in deduction. Unlike deductive research, induction seeks to establish "consistencies, patterns and meanings through a process of collecting data" (Gray, 2014, p. 19).

3.3.2 Qualitative approach

The research methods in social sciences are generally divided into two principal categories: quantitative and qualitative methods (Muijs, 2004). In social research one of these two approaches can be selected to conduct an investigation. Alternatively, mixed methods, which combines quantitative and qualitative approaches, can be adopted. The choice of methods depends on the type of data a researcher needs to gather to test a hypothesis or to answer a research question (Bryman, 2016; Matthews & Ross, 2010).

The selected methods for this research are qualitative methods. Qualitative research centres its attention on words rather than on numerical data. It is appropriate to conduct qualitative research when a complex understanding of an issue is required, that is to say, when interactions among people are wanted to be captured, and context or settings in which interactions take place are wanted to be explored (Creswell, 2013). The type of findings that can be gained from qualitative methods are therefore useful to address the research problem of this study.

Quantitative methods, on the other hand, are not useful for the purpose of this research since this study does not intend to explain a phenomenon by collecting numerical data and the objective of this research is not to test a theory, i.e. a deductive approach. Overall, quantitative methods are often more suitable to studies adopting practices related to positivism. The following section discusses qualitative approaches in more detail.

Qualitative research generally centres its attention in words, rather than on quantification when collecting and analysing data (Bryman, 2016). Unlike quantitative research in which existing theory is tested through research, in qualitative methods, theory emerges out of research. This type of research rejects the practices adopted in the natural scientific model (in particular, of positivism) (Bryman, 2016). Qualitative research takes an inductive approach to the association between theory and research, and prioritises the development of theories (Bryman, 2016).

Qualitative methods pay attention to the interpretations that individuals make about the social world (Bryman, 2016) and are mainly concerned with accounts and stories including beliefs and opinions, feelings and subjective understandings (Matthews & Ross, 2010). For this reason, in the data collection process, words or expressions of the research participants are gathered rather than numerical data (Matthews & Ross, 2010). In qualitative research, generally researchers do not limit their data collection to a single source; rather they seek to gather different forms of data by interviewing people, observing behaviours, and examining documents; commonly the data collection process occurs in the places where participants experience the issue or problem that is being investigated (Creswell, 2013).

In qualitative research, researchers do not intend to establish cause-and-effect links among factors; rather they have the objective of identifying complex interactions and relations of factors in any situation (Creswell, 2013). Reporting multiple perspectives and identifying different factors involved in a situation becomes crucial to build a rich, solid, and detailed picture of the phenomenon that is being investigated (Creswell, 2013). The central idea of qualitative research is to learn about the issue or problem under investigation from the participants and adopt the most suitable strategies to obtain information. It is therefore crucial for the researcher to be aware that the research design is emergent, that is to say, that the initial research plan may change once the researcher steps into the field and starts gathering data: individuals studied, research questions, forms of data collection and visited sites may be modified (Creswell, 2013). The researcher also needs to be aware that qualitative research requires spending many hours in the field and engaging in an arduous data analysis process.

3.4 Research strategy and design

A research strategy consists of designing a plan to conduct the research (Matthews & Ross, 2010). According to Saunders (2015), a research strategy is a plan for how a researcher intends to address a question posed. In a research strategy, harmony and a coherent connection between the research philosophy and the data collection and analysis methods are expected (Denzin & Lincoln, 2011).

3.4.1 Data Journeys approach

My initial interests to conduct this research revolved around the social dynamics shaping health data flows and reuse practices. The Data Journeys approach developed by Bates et al. (2016) and its underlying philosophy seemed the most suitable to conduct this study. It was selected because it is aligned with key assumptions in the CDS field, attends to interpretivist considerations when examining reality, and acknowledges the socio-material constitution of data and data flows.

As explained in section 2.5 of this work, the data journeys approach (Bates et al., 2016) was developed as a methodology to examine the socio-material composition of data objects and flows as data moves across diverse cultures and sites where different data practices are performed, from their generation point through to reuse in various sectors. It aims to uncover the specific ways in which different sociocultural values and material elements coalesce over time to generate the socio-material conditions that shape data practices (e.g. data generation, processing, and dissemination), and consequently impact upon the use of data and their movement between infrastructures (Bates et al., 2016).

The DJ approach to research design is described below, including its main characteristics and how this methodology is applied. Theoretical aspects are not discussed as they were addressed in the literature review of this work, section 2.5.

An approach influenced by mobile ethnography ideas

The DJ approach was developed drawing on the ideas of scholars who in the past have adopted the process of journeying as a method, such as the work of Raymond Williams (1958) and research that has adopted a 'mobile ethnography' approach. Bates et al.'s (2016) approach is one in which the researcher moves across different sites of data practice, which are all those places where data are generated, processed, disseminated, and used (research groups'

headquarters, networks, archives, repositories), with the intention of following data on their 'journey'. The aim is not only tracing their movement, but also adopting the role of a traveller, stopping off on the journey to pay attention to the surroundings and to soak up the culture of diverse sites of data practice.

The approach requires the researcher to follow data as they travel across interlinked projects and institutions inside and across infrastructures (Bates et al., 2016). It pays attention to how and why data move from one point to another and issues related to, e.g., obstructed movement or lack of movement.

In general, the data journeys methodology consists of (Bates et al., 2016):

- Identifying important sites of data production and conducting an initial mapping of key data journeys;
- Making initial enquiries regarding access to research sites and then deciding on which data journeys to focus;
- Detailed mapping of the selected data journeys between pertinent institutions, projects, datasets and people;
- Following these data journeys, identifying key informants at each site of practice using desk research and chain-referral or snowball sampling, and important documents;
- Generation of primary data and collection of secondary data aimed at understanding in what ways sociocultural values interact with existing material conditions to act as drivers or constraints to move data between different sites:
    - In-depth interviews with data practitioners and other relevant individuals
    - Reflective field notes at different sites of practice
    - Documentary analysis of relevant policies and legislation
- Final mapping of selected data journeys based on primary and secondary data;
- Analysis of data collected to address the question of how sociocultural values interact with existing material conditions to generate drivers and frictions that work together to produce the material forms of data journeys in the context of reuse of NHS patients' personal data for research purposes.

Application of Data Journeys in this research project

Data Journeys (Bates et al., 2016) was selected as a valuable approach to uncover:

1. How patient data are generated, processed and used, and move between various sites of data practice linked by the flow of patient data across space and time;
2. How sociocultural values and material elements assemble to shape and provide justification to these flows and practices; and
3. How in combination these sociocultural values and material factors contribute to shaping data flows.

The data journeys methodology was used with some adaptations to explore the journeys of selected personal patient data produced by the NHS, from the initial generation and collection through to reuse for secondary purposes in universities. This study centres its attention only on exploring data flows and not on the constitution of data objects as Bates et al. did in their research on weather data. Furthermore, it does not integrate an oral history element to the interviews conducted. Another adaptation was that in this study contextual field notes were taken to build up an understanding of the context, but field observations and digital ethnography were not incorporated.

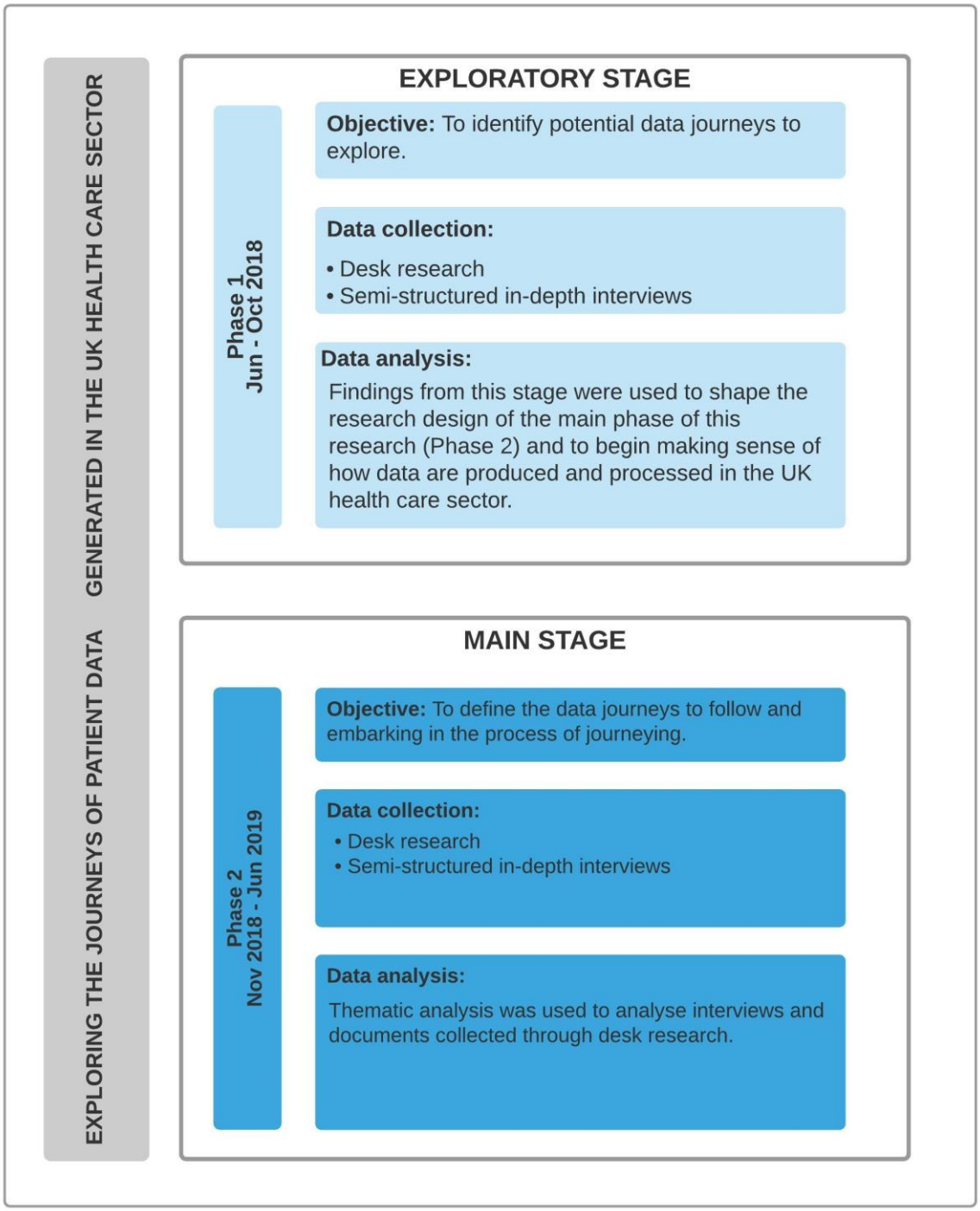
3.4.2 Research design

Following the Data Journeys first step above, I initially planned to follow the journeys of patient data flowing to two different types of sites: universities and pharmaceutical companies. In order to do this, Iintended to conduct interviews with key informants at these two types of sites as well as informants at data intermediary sites (organisations that process and provide access to patient data to both universities and pharmaceutical companies).

As discussed further in section 3.5.1.2, I was able to recruit key informants at universities; however, significant recruitment difficulties were experienced with pharmaceutical companies and data intermediary sites. Due to the challenges experienced in recruiting key informants at pharmaceutical companies and data intermediary sites, it was not possible to explore such journeys and a decision was made to focus the research design on reuse of patient data for research in universities.

Once the research focus was decided, the study was conducted in two phases. The objective of Phase 1 was to identify a number of potential data journeys to explore. Phase 2 consisted of further refining the specific data journeys to follow and embarking on following the data as they journeyed through different organisations or projects. Figure 1 shows an overview of the data collection conducted in the two phases of this study.

Figure 1. Research design



3.5 Data collection

This section describes the data collection conducted in this study.

3.5.1 Semi-structured, in-depth interviews

In the two phases of this study, semi-structured in-depth interviews were conducted. The utilisation of semi-structured interviews enables the researcher to manage in a better way the interview topic than in a unstructured interview (Given, 2008). This strategy provides flexibility to the interviewer, permits them to adapt the interview to the respondent, and allows the interviewees to give answers with more freedom (Miller & Brewer, 2003).

In-depth interviews, which are also referred to as non-directive interviews, active interviews, focused interviews, and open-ended interviews (Curtis & Curtis, 2011b) are among the most common methods used to collect data in qualitative studies (Given, 2008). The objective of in-depth interviews is to gain a deep understanding of a topic, and they are suited to information-rich cases (Curtis & Curtis, 2011b; Patton, 1990). In these interviews, participants are motivated to talk in depth about events, beliefs and behaviours related to the topic or issue under investigation (Given, 2008; Saunders et al., 2015). In in-depth interviews, the quality of the interaction between the interviewer and research participant is paramount (Curtis & Curtis, 2011a).

In this type of interview predefined, focused questions are not used (Given, 2008). The researcher does not need to develop an extensive list of questions; nevertheless they must have a clear idea about the topics that are desired to be explored (Saunders et al., 2015). The researcher needs to be aware of the main domains of experiences that are possible to discuss by the participant and be capable to probe in which way these relate to the issue under investigation (Given, 2008). In-depth interviews are developed using schedules that include questions or topics that are flexible, and are not highly structured in terms of the questions asked, because it is considered that this could restrict the depth of data collected (Curtis & Curtis, 2011b). An in-depth interview may include a mix of closed questions (to ascertain interviewees' demographics) and open-ended questions, or even no formal questions at all.

A description of the interview topics is provided in section 3.5.1.3.

3.5.1.1 Phase 1 Sampling and recruitment

In this phase, I conducted semi-structured in-depth interviews with experts who are familiar with the use and processing of patient data within the health and care sector. In order to recruit

suitable candidates to be interviewed, I conducted initial desk research, and as a result of this it was determined that it would be useful to interview experts from different backgrounds (e.g. senior academics, senior members of key organisations such as NHS Digital) who share a common interest on the uses of patient data.

Potential participants were identified and invited to participate in the research via email. A total of four interviews were conducted in Phase 1. Demographic characteristics of participants are presented below:

Table 1. Demographic characteristics of interviewees. Phase 1

Participant's background	Position	Gender
Academic	Professor of Health Informatics	M
Not-for-profit organisation	Senior data manager	M
Public organisation	Senior staff at NHS Digital	M
Public organisation	Senior staff at NHS Digital	F

3.5.1.2 Phase 2 Sampling and Recruitment

In this phase, I conducted semi-structured in-depth interviews with team members of research groups based at UK universities who are familiar with the use and processing of personal patient data.

The recruitment process of Phase 2 was conducted in two different ways. Initially, I approached potential interviewees via email; however this strategy was only partially successful, as I only managed to recruit three interviewees through this. After the low success of participant recruitment via email, I adopted an alternative strategy, which consisted of attending conferences, presentations or other relevant events where potential participants could be approached. Using this strategy I managed to recruit a larger number of interviewees.

Examples of the events attended to recruit participants are presented below:

1. Meet the Stemettes @ NHS: Leeds Edition, 18th October 2018.
2. The future of digital technology in the NHS - Big Data and AI, efficiency and outcomes, and addressing concerns, London, 23th November 2018.
3. HDR UK Digital Innovation Hub, Nottingham, 14th March 2019.
4. Health Data Research UK London seminar, 18th March 2019.
5. UK Biobank Scientific Conference, London, 19th June 2019.

According to Curtis & Curtis (2011b), the sampling strategy for in-depth interviews must be purposeful, and generally speaking the number of participants in an interview-based study tends to be small. They also argue that it is important to take into account that it is difficult to determine the size of an interview sample ahead of time because it is necessary to revise the data as it is gathered to decide when to stop data collection (Curtis & Curtis, 2011b). There are a number of approaches that can be adopted to determine the sample size in research of qualitative nature. For example, Klenke (2008) suggests using a sample size of two to twenty-five participants. In a similar fashion, Creswell (2007) recommends a sample size of five to twenty-five individuals. Other scholars recommend the researcher stop doing interviews only after achieving data saturation, that is to say, that new themes cease appearing (Bowen, 2008; Mason, 2010). In this phase, I exhausted all leads on every one of the data journeys and conducted as many interviews as possible at each site to get a deep enough understanding of the issues explored. In addition to this, as data was analysed I got to a point where similar points of view and perceptions were identified in different interview transcripts. A total of 18 interviews were conducted. The demographic characteristics of the sample are presented in the table below:

Table 2. Demographic characteristics of interviewees. Phase 2

Site	Position	Gender	Participant ID
SITE 1	Researcher	F	S1-YR-1
	Researcher	M	S1-LR-2
	Researcher	M	S1-NR-3
SITE 2	Researcher	F	S2-FA-1
SITE 3	Researcher	F	S3-EN-1
SITE 4	Senior researcher	F	S4-AS-1
	Researcher	M	S4-LS-2
	Researcher	M	S4-NS-3
	Researcher	M	S4-SS-4
	Researcher	F	S4-YS-5
SITE 5	Senior researcher	M	S5-IN-1
SITE 6	Researcher	M	S6-NM-1
	Researcher	F	S6-AM-2
	Researcher	F	S6-HM-3
SITE 7	Researcher	M	S7-GM-1
	Data manager	F	S7-GN-2
SITE 8	Researcher	F	S8-RN-1
	Researcher	F	S8-RA-2

As mentioned in 3.4.3, significant challenges were experienced when approaching key informants in pharmaceutical companies and data intermediary sites.

In an effort to recruit informants in pharmaceutical companies, I explored the Clinical Practice Research Datalink (CPRD) website. CPRD is a not-for-profit research service of the UK government that provides access to anonymised patient data collected from a large number of GP practices across the UK for observational and interventional public health research (Clinical Practice Research Database [CPRD], 2018). Through this exploration it was possible to identify general information about a number of pharmaceutical companies that were granted access to patient data to conduct research from April to November 2018. The information gathered included the research name of the company accessing data, names of research projects and brief summaries.

After identifying a number of potential data journeys to explore through CPRD, I proceeded to search for additional information on the official websites of the identified pharmaceutical companies to gather contact details of key informants and to begin making sense of how companies use patient data for research purposes. Whereas exploring the official websites of university-based research teams was a useful strategy to gather relevant information about some aspects of data flows, this was not the case in the websites of private companies. University-based research teams' websites tend to offer plenty of information in relation to their research projects (e.g. research project overview, data used, data collection and analysis methods, key information about team members, including names, background, research interests, roles, and contact details). In contrast, pharmaceutical companies' websites tend to provide very limited information, and most of it has the purpose of showcasing their products and services.

Through the exploration of companies' websites, I was only able to gather general information about the research interests and main topics explored, broad explanations of data analysis methods used across the company but not in specific projects, and brief individual profiles of some research team members, but no contact details of potential key informants.

Given that the information obtained from the websites explored was very scarce and not useful for the purposes of this study, I decided to explore and pursue an alternative avenue to establish contact with potential key informants. The strategy consisted of approaching the companies through the contact form provided on their websites. This strategy was unsuccessful as none of the companies approached responded.

In addition to the challenges faced to approach key informants at pharmaceutical companies, I also experienced difficulties in establishing contact with informants at data intermediary sites,

which are organisations that process and provide access to patient data for research purposes to both universities and pharmaceutical companies.

I intended to conduct interviews with informants at two key data intermediary sites, The Health Improvement Network (THIN) and Clinical Practice Research Datalink (CPRD); however it was extremely difficult to establish contact with potential informants in both sites. These organisations were approached via their contact form available at their official websites. I sent detailed information about the project and enquired about the possibility of conducting interviews with practitioners and senior staff. THIN was approached on three different occasions; however they did not respond to any of these attempts.

CPRD promptly responded to the researcher enquiry, informing that the request was going to be considered. Shortly after that, I was contacted by a senior member of CPRD, who asked me to provide the list of interview questions for their consideration. Whereas this was an unusual request, I agreed to this and shared a list of potential questions to discuss during the interview. In response to this, I was informed that it was not going to be possible to conduct face-to-face interviews. The reason that the person gave for not participating in the interview was the following: “Given the nature of your questions and CPRD operations, interviews with individual staff members will not be helpful or meaningful for your research”. While the face-to-face interview was declined, the contact at CPRD offered to provide “corporate responses” to some of the questions that the researcher sent. I accepted the offer and four weeks later, the "corporate responses" from the organisation were sent. Only brief responses to selected questions were provided; however the majority of them were not answered as according to the informant these questions were not really relevant for them. These responses were integrated to the documents thematically analysed in Phase 2 of this research and were also useful to answer factual questions to inform the descriptions of data journeys examined in this study.

Despite it not being possible to establish contact with key informants and obtain access to these sites, the exercise of seeking access to them revealed features about them, and showed key points of comparison. The challenges experienced in getting access to these sites were useful to uncover an important aspect of the Data Journeys methodology: the full potential of it can only be realised when stakeholders involved show willingness to participate and are transparent; however these challenges can also be used to point to black boxes in data sharing infrastructures. For example, I identified that whereas university-based research teams are often approachable and contextual relevant information about them and their projects can be obtained through desk research, this is not the case with pharmaceutical companies. These

organisations appeared to be less willing to discuss their work with external stakeholders. I also identified that official websites of these organisations only offer very limited insights concerning their data practices. Throughout the duration of this study, I became aware of controversial cases concerning the circulation of data towards particular pharmaceutical companies, which have appeared not to have taken into account patient privacy and agency. However these cases were not disclosed by the companies or the NHS; rather they were unveiled by journalistic investigations (Powles & Hodson, 2017).

3.5.1.3 Conducting the interviews

Van Teijilingen & Hundley (2002) observe that in qualitative research, a subsequent interview in a series is commonly a better interview than those conducted before. This is because it is very likely that the researcher has obtained valuable insights from previous interviews and used them to refine the interview guide and improve their practice. In this study, I refined and strengthened my practice through the data collection process. The questions' structure and order were polished to encourage and foster more detailed explanations from the interviewees. In all of the interviews conducted in phases 1 and 2 of this study, I first introducedmyself. Then I reminded participants of the aim of the research and the interview's objective. Consent from all participants to participate and to record the interviews was obtained. All interviews were audio-recorded and transcribed.

Interviews conducted in Phase 1

I conducted four face-to-face interviews with experts from different backgrounds who are familiar with the use of patient data for research purposes outside the health and care sector. These interviews were conducted between June and October 2018. The average duration was 45 minutes. The following topics were covered in these interviews:

1) Participants' professional background, current position and area of expertise.
2) Examples of data flows in the healthcare sector in which they perceived: a) patient data flows with few frictions between different people and organisations, and b) there are too many barriers restricting the movement of patient data between different people or organisations.
3) Main challenges to overcome when sharing patient data for purposes other than the direct care of patients.

4) Participants' views concerning the role of existing relevant guidance and regulations within data flows in the healthcare sector.

Interviews conducted in Phase 2

During Phase 2 of this research, 18 in-depth interviews were conducted with team members of different research projects working at UK-based universities who are familiar with the use and processing of personal patient data for secondary uses. With consent from the participants, I spent short periods of time at the sites of data practice visited before and after conducting the interviews to take reflective contextual field notes to build up an understanding of the working environment and dynamics. Interviews were conducted between November 2018 and June 2019. Aspects covered in these interviews:

The role of the interviewee and the organisation:

1) Participants' professional background, current position, key responsibilities
2) Participants' work team characteristics (e.g. structure, team members, main responsibilities)
3) Participants' perception regarding their own team's culture and wider culture of the organisation where they work

The work with data:

1) Types of data collected and reused
2) Examples of recent research projects conducted by their team reusing patient data and details about them. For example:
    - How was data collected?
    - Who was involved?
    - What are the main challenges encountered when working with these data? How did they make sense of and analyse data?
3) The practices around access to patient data, including data access request processes, interactions with data providers and data producers, etc.

The reuse of patient data

1) Benefits and risks concerning the reuse of patient data
2) Challenges and drivers perceived concerning the reuse of patient data

3) Views concerning regulations that dictate how and under which circumstances data can be shared and to whom
4) Perceptions concerning the reuse of patient data by different organisations and for different purposes (e.g. charities, pharmaceutical companies, university-based research teams)
5) Perceptions concerning the culture of the above mentioned organisations and people working for them

3.5.2 Desk research

Desk research was undertaken in both phases of the study to gather primary documents. This section presents details about how this process was conducted.

In order to define the source for documents, the context of the topic of study and the research question should be considered (Frey, 2018). In this study, inclusionary criteria were helpful to ensure that the document selection was conducted in a systematic way and to reduce irrelevant data collection. On the other hand, exclusionary criteria were helpful to narrow the number of potential documents down to the final sample. Details about the inclusionary and exclusionary criteria applied are provided in sections 3.6.3.1 and 3.6.3.2.

In phases 1 and 2, I collected outputs produced by key stakeholders in the UK healthcare landscape to develop a good understanding of key data sharing policies, legislation, and discourses in key documents produced by stakeholders within the UK healthcare landscape. In addition to this during Phase 2, I collected documents produced by university-based research teams with two objectives. First, to gather details to enrich the descriptive accounts of the journeys of patient data selected for this study and the material infrastructures that support such flows, which were mainly informed by interview data. Second, to get a deeper understanding of sociocultural values and norms that play a significant part in shaping data practices within university-based research teams.

3.5.2.1 Sampling procedure of documents produced by key stakeholders in the UK healthcare landscape

Document sources

One of the sources of primary data collected for this research was the website of UK legislation (legislation.gov.uk). Official websites of data creators, data providers, and funding bodies were

also explored to gather additional relevant documents. In addition to this, during Phase 2 I collected relevant documents disseminated by organisers and presenters at the public events she attended to recruit key informants.

Sampling procedure

In a preliminary search within official websites, a large number of documents were identified, therefore I defined inclusionary and exclusionary criteria. I was interested in a current representation of the construct of study, therefore I choose to limit the time period for the document creation to the last five years. However, I also acknowledged that some documents produced in previous years were relevant in the context of this study, therefore exceptions were made and key documents were included due to their relevance despite not meeting the document age criteria. Examples of documents older than five years that were included are: The Health and Social Care Act 2012 (Department of Health and Social Care, 2012), and the Caldicott review 'Information: To share or not to share?' (Department of Health, 2013a). I decided to exclude from the final sample those documents that were considered less relevant or appropriate to help developing a deeper understanding of the UK health data landscape.

In the case of documents collected at public events, I applied an inclusionary criteria based on the type of document. I decided only to include event reports, presentations slides and presentation transcripts and to exclude outputs such as posters, infographics, and programmes. I also decided to limit the number of outputs to five per event.

Examples of documents collected from official websites:

1. Health and Social Care Information Centre is now called NHS Digital (Department of Health and Social Care, 2016).
2. Information to share or not to share? The Information Governance review (Department of Health, 2013a).
3. Independent Group Advising (NHS Digital) on the Release of Data (IGARD): Terms of Reference (Carrigan & Williams, 2018).
4. Initiatives in data science research (Medical Research Council, n.d.-b)
5. Launch of the Clinical Practice Research Datalink (Department of Health and Social Care, 2011).
6. National data opt-out guidance for researchers (NHS Digital, 2019c).

7. National Health Service Act 2006 (Department of Health and Social Care, 2006).
8. Health and Social Care Act 2012 (Department of Health and Social Care, 2012).
9. The value of using data (Medical Research Council, n.d.-c).
10. Understanding the health and care information we collect (NHS Digital, 2019b).

Examples of documents collected from public events:

1. Presentations slides and transcripts of all speakers from the seminar 'The future of digital technology in the NHS - Big Data and AI, efficiency and outcomes, and addressing concerns', London, 23th November 2018 (Westminster Health Forum, 2018).
2. Programme and presentation slides of keynote speakers from the engagement event 'HDR UK Digital Innovation Hub', Nottingham, 14th March 2019 (HDR UK, 2019).
3. Presentation of all speakers from the annual event 'UK Biobank Scientific Conference', London, 19th June 2019.

3.5.2.2 Sampling procedure of documents produced by university researchers

Documents produced by university-based research teams were collected in two different ways. I gathered relevant documents through an exploration of the official of university-based research teams that participated in this research. In addition to this, interviewees provided printed copies of documents or directed me towards what they considered as key outputs in their websites. I only included documents that were produced in the last two years.

Examples of key documents collected from sites of data reuse include:

- Data sharing agreements
- Data collection charts and diagrams
- Annual reports
- Project outcome briefings
- Research project outlines
- Research reports
- Blog posts

3.6 Data analysis strategy

In qualitative research, the analysis process allows the researcher to make sense of data and generate meaning from them (Merriam, 2009). This section describes how data analysis of this project was conducted.

I transcribed, listened carefully to all interviews and took notes.

3.6.1 Phase 1 – initial mapping

In phase 1, the analytic approach was informed by the Data Journeys methodology (Bates et al., 2016). Interview data from this phase was used specifically to identify potential data journeys to follow and begin making sense of how data produced in the healthcare sector flows to different sites of data reuse. The analysis process therefore involved mapping techniques: in this process, the researcher created visual representations of the journeys in two ways, using post-it notes on paper or using the web-based drawing platform Lucidchart.

These initial steps in the data journeys approach were used to shape the research design for phase two and to begin making sense of how data are produced and processed in the UK healthcare sector. The analysis of data collected during the first stage of this study allowed me to obtain a deeper understanding of the variety of patient data flows that exist in the UK healthcare sector and the social actors involved in them. It also allowed me to identify key issues and concerns related to the uses and sharing of patient data in the UK healthcare sector.

3.6.2 Phase 2 - Thematic Analysis

Thematic analysis was applied to conduct the analysis of the 18 interviews conducted in Phase 2, and documents collected through desk research in the two phases. As mentioned in section 3.6.3, some documents gathered from sites of data reuse such as technical reports or data sharing agreements were used to enrich the descriptive accounts of the data journeys explored. These types of documents were not thematically analysed, rather they were treated in a different way. These types of documents were only interrogated to answer factual questions in relation to the flows of data (e.g. what type of data was collected, when the sharing agreement was approved, data linkage processed used) and contribute to the mapping process.

In this research, the thematic analysis approach developed by Braun & Clarke (2006) was followed. This process, which is described in detail in section 3.6.2.1-6 was informed by ideas

and recommendations proposed by Fairclough & Fairclough's (2012) approach to critically analyse discourses.

Thematic analysis is a useful strategy for identifying, making sense and reporting patterns in datasets (Braun & Clarke, 2006). Through the application of thematic analysis, the researcher is able to organise data in a consistent way and develop rich and detailed descriptions (Braun & Clarke, 2006). Thematic analysis can be useful to analyse various types of data, such as field notes from ethnography, interview and focus group transcripts, historical or site documents, drawings, sets of images, maps, digital video files and audio files (Bryne, 2017; Mills et al., 2010). In addition to this, it has been considered the most appropriate method to analyse data in studies that have the objective of discovering using interpretations (Alhojailan, 2012).

Fairclough & Fairclough (2012) invite the researcher to pay attention to different elements of an argument, which are:

1. Claims that are made to answer "What should I (we) do?" to address a specific problem or challenge under certain circumstances.
2. Circumstantial premises, that is to say, how current states of affairs are pictured, understood, or 'problematised'.
3. Goal premises, which are feasible and attractive possible solutions.
4. Value premises, which are underlying values and concerns.
5. Means-goals premises, which are premises of a conditional form that establish that if a specific action or a series of actions are performed, we would be able to overcome an existing problem or issue and achieve a desirable and attractive reality.

Throughout the data analysis process, I carefully examined 'persuasive definitions' and 'emotive terms'. These are very important because this suggests that "arguers are oriented towards particular resolutions or ideas and not others" (Fairclough & Fairclough, 2012, p. 92). In this analysis, as Fairclough & Fairclough recommend, persuasive definitions and emotive terms were understood as "arguments, with a burden of proof attached" (Fairclough & Fairclough, 2012, p. 93) which should be critically questioned.

I also paid attention to metaphors or frames because they play a key role in determining how people conceptualise their goals and circumstances, and therefore how they act. Analysing metaphors can be helpful to understand how re-framing a situation operates within the plan of action presented, how it provides people with reasons or justifications for actions and how it

fits within a specific action plan (Fairclough & Fairclough, 2012). Following the suggestion given by Fairclough and Fairclough, representations were analysed as parts of premises of arguments rather than as elements in isolation.

Fairclough & Fairclough's (2012) approach places great importance on imaginaries as according to them, they can motivate and inspire actions. It also suggests that imaginaries can have a performative power when they are collectively conceived as representations of actual, not only possible states of affairs. In this analysis, I acknowledged that discourses presented in institutional documents exist within the context of their creation, and therefore their creation is influenced by social, economic, political, and cultural factors.

3.6.2.1 Familiarisation with the data

Familiarisation with data is the first step in thematic analysis according to Braun & Clarke (2006). This step consists of transcribing data, reading and re-reading data, and making notes capturing initial ideas. In this phase, I began by transcribing interviews previously recorded into a written form using Microsoft Word, assisted by a transcription foot pedal. Preparing and organising collected data allow me to increase my understanding of the data and begin to develop a meaningful analytic approach. After data transcription, interviews were re-read, nascent patterns and themes were identified, and a number of preliminary codes were generated.

3.6.2.2 Generation of initial codes

After familiarisation with the data, it is necessary to generate initial codes. This stage consists of systematically coding features that draw the attention of the researcher across the data collected (Braun & Clarke, 2006). In this phase, the generation of codes was conducted combining two different approaches: manual coding and coding assisted by NVivo, qualitative data analysis software. The manual coding offered me the possibility to connect ideas by spreading out pieces of paper on the floor and this helped to make sense of a large puzzle with many small pieces. On the other hand, NVivo was a useful tool for handling data gathered from different sources, such as documents and interviews (Jupp, 2006).

The process began with open coding. Additionally, some key concepts were identified through the process of reviewing the literature, and these theory-related concepts were used by the researchers as springboards for themes (Ryan & Bernard, 2003). In the initial stage of the analysis and based on three selected interviews, fifty-four codes were generated. Similar codes

were then divided into different groups. Later on, the codes were reviewed with the objective of identifying and eliminating repetitions. The iterative and continuous coding procedure allowed me to develop significant and relevant themes. During this process, I took notes to build up associations between diverse terms and ideas. After the analysis of the complete dataset, new codes were identified. In some instances the new codes were more precise and clear than previous ones. An extract of the list of the codes generated is presented below:

- The perceived benefits of sharing data for secondary uses
- Interest in improving patient outcomes
- Patients' involvement is key
- A cultural shift around data use and reuse is needed
- Scrutiny of public
- Trust and confidence from the population
- Efforts to prevent breaches are needed
- Data is never completely anonymised
- Bureaucratic processes associated with data providers
- Ethical challenges
- Economic barriers / funding
- Technical barriers
- Governance processes
- Lack of understanding of legislation, regulations
- People in the healthcare sector are reluctant to share data
- Care.data backlash

3.6.2.3 Searching for themes

In this stage, the analysis is re-focused at a wider level, paying attention to themes instead of codes. Searching for themes requires organising the codes into prospective themes and gathering all meaningful and significant coded extracts of data within the themes generated.

In the process of searching for themes, I was also mindful of Ryan and Bernard's (2003) suggestion of paying attention to:

- **Repetitions:** topics that emerge constantly
- **Similarities and differences:** paying attention to how people discuss topics in diverse ways
- **Linguistic connectors:** analysing the utilisation of connectors (e.g. 'therefore' or 'because', 'hence') because words like those may point out causal connections perceived by participants
- **Missing data:** paying attention to what is absent from the data, considering for example what participants do not mention in their responses or what is not said in documents

During this stage, I formed overarching themes by combining or separating different codes. In order to discover connections between the codes and themes generated, I constantly compared and reflected on the emerging themes. The last step of this phase involved labelling extracts of data with initial themes and subthemes. A list of the main themes generated in this stage is presented below:

Characteristics of health data

1. Messy
2. Inaccurate
3. Incomplete
4. Data gaps
5. Miscoded
6. Stored in different systems

Characteristics of health data flows

1. Underused
2. Travel in many different directions
3. Unclear for data creators
4. Confusing for research teams
5. Unclear for the public

The perceived value of data

1. Useful to advance careers of researchers
2. Can improve health outcomes
3. Useful to solve health issues
4. Efficient and cost effective

Main concerns around data sharing

1. A cultural shift around data use and reuse is needed
2. Scrutiny from the public
3. Trust and confidence from the population
4. Data needs to be exploited

Enablers of data movement

1. Affordances of external tools and technologies
2. Safe and secure ways of sharing data are available

Sources of data friction

1. Access requirements
2. Organisational priorities
3. Organisational structures
4. Reluctancy from people in the healthcare sector
5. Data scandals and Care.data backlash

3.6.2.4 Reviewing themes

The objective of this phase is to reflect on the validity of each theme in relation to the whole dataset, and to verify that the thematic scheme presents an accurate representation of the meanings within the dataset (Braun & Clarke, 2006). Thus, I refined and reworked the list of candidate themes previously created. This process involved breaking down some of the themes into independent themes, and combining others into a single theme. A list of the reviewed themes is presented below:

Drivers for the flow of data

1. Interest in advancing health research
2. The perceived potential of routinely collected patient data
3. Technological advances
4. Interest in advancing academic careers

Sources of data friction

5. Data sharing infrastructures and management
6. Regulatory frameworks
7. Sociocultural factors

Overcoming data frictions

8. Patient and public opinion to mitigate risk aversion
9. Patient and public support as a source of legitimacy
10. Patient groups as an opportunity to engage in meaningful conversations

3.6.2.5 Defining and naming themes

After reaching a satisfactory decision in regards to the themes created, I further refined her analysis and captured the soul of each theme by clearly defining the aspects of the data they represented (Braun & Clarke, 2006, p. 92). In some cases where a theme included too many codes, I followed Braun & Clarke's (2006) advice of going back to the units of data that were gathered within all themes to verify that their names were not too complicated; themes must transmit to the reader a clear idea of their meaning.

3.6.2.6 Producing the report

This final stage consists of writing the findings report in order to offer a clear, consistent, reasoned and engaging narration of the story that was developed through the analysis of data (Braun & Clarke 2006). Chapters 4 to 7 present the overall findings of this study, and Chapter 8 discusses the wider theoretical perspective of this research.

3.7 Challenges of tracing and visualising health data journeys

This section presents a reflection about the challenges of tracing and visualising health data journeys. First I focus on the key challenges I experienced while tracing data journeys and later on I reflected on the difficulties of visualising them.

I encountered three key challenges in the process of tracing data journeys. The first one is that it was not possible for me to gain access to sites of data reuse controlled by private sector actors. As mentioned in the methods chapter of this thesis, I initially proposed to follow journeys of patient data in the academic context and in the private sector (pharmaceutical companies), I was able to trace the journeys of patient data flowing towards universities, however I could not access sites of data reuse controlled by private actors mainly due to issues of transparency. Through desk research I was able to gather some limited information about potential journeys to follow as they flow towards pharmaceutical companies, for this I mainly relied on public information available in the website of Clinical Practice Research Datalink (CPRD), a not-for-profit research service of the UK government that provides access to anonymised patient data collected from a lage number of GP practices across the UK for health research, however I could not progress in tracing these journeys further than that. After this initial step I approached a number of pharmaceutical companies but my attempts were unsuccessful as none of them replied to my communications. In addition to this, I also attempted to gather relevant information through other sources such as the official websites of these sites of data reuse but the information found was very scarce and not useful for the purposes of this research. As explained elsewhere in this thesis, this challenge served to uncover an important feature of the Data Journeys methodology, which is that the full potential of it can only be realised when stakeholders involve show willingness to participate and are transparent. Despite it was not possible to access this site, the exercise of seeking access to these sites revealed interesting features of these sites, and show interesting points of comparison.

The second challenge was that it can be extremely difficult to make sense of the different paths that data take. This challenge was not a surprise, the NHS itself has acknowledged that flows of data generated in the heathtcare sector are verly complex and messy (Centre for Data Ethics and Innovation, 2020; Harwich & Laycock, 2018) and that understanding where data are and the processes in place to access them is not an easy task. While interviews conducted with key informants were useful to understand to some extent some aspects of the processes required in

order to obtain access to patient data, the technical arrangements that need to be in place prior to accessing data, and the roles of different stakeholders that are involved at each stage of these data journeys, to produce a more accurate picture of the data journeys traced it was also necessary to spend a significant amount of time conducting desk research to gather relevant documents from different sources.

The third challenge was that due to time constraints, it was not possible to trace some of the journeys identified as candidates to be followed in the first stage of this research. Due to the sensitive nature of patient data, before accessing them, university-based researchers need to obtain an ethical approval for their projects and this process can take several months. In the case of this research, some of the journeys that I attempted to follow did not starton time for me to follow them because researchers were awaiting to obtain approval to commence their projects. Researchers interested in tracing journeys of health data in the future therefore should bear in mind that sometimes data can be stalled for quite a long time before they can begin travel towards sites of data reuse; this can set a challenging scenario if limited amount of time is available for tracing journeys. In the case of my research, while it would have been interesting to trace additional data journeys, I could not spend more time doing field work as I was constrained by the limited time that I had available for conducting my fieldwork.

Creating visual representations of data journeys also posed a number of challenges for me. The process of developing visuals allowed me to learn that they are useful to tell the story about the path that data follow as they travel between different sites of practice (how they move from one point to another), but less effective to depict the sociomaterial factors that shape data flows. Despite several attempts I could not successfully integrate to the visuals key frictions identified in this research (e.g. frustrations experienced by researchers when being denied access to patient data, slow down in the movement of patient data due to infrastructural barrriers). In the same way it was equally complicated to depict drivers for data to move and how they interact between each other to shape data journeys (eg. how the enthusiasm and excitement of researchers for conducting research with patient data combined with the provision of material resources by funders and other key stakehoders has helped data to flow to the hands of university based researchers). It was also complicated to decide what to visualise. While visuals can help with explaining tricky concepts and processes that take place within systems (Annan-Callcott, 2021) (e.g. how data linkage, aggregation and deidentification are conducted) I realised that trying to add a visual representation of these processes to the visualisations would

have make them overwhelming or overcomplicated and might have generated confusion for those engaging with these visuals.

Finally, while I made an effort to produce effective visuals, this tak was challenging for me because I am not a trained designer. Because of this, I had to spend a long time exploring different available tools for producing visuals, becoming familiar with the selected software to produce the visualisations, and working on several drafts before producing a satisfactory final version. Based on this experience, I believe that researchers interested in producing effective and engaging visuals would benefit from integrating to their team a professional designer, if this is not a suitable alternative due to limited resources, then researchers should keep in mind that producing visualisations is a complex and time consuming endeavour and take this into account when designing and planning their research.

3.8 Research quality

It is important that researchers demonstrate that their study is high quality, that is to say that the findings presented are valuable and deserving of attention (Lincoln & Guba, 1985). The criteria to assess quality is not the same for qualitative and quantitative research because these two approaches are different in nature (Bryman, 2016). In quantitative research, key factors that are markers of high quality include validity and reliability (Bryman, 2016). In qualitative research, there is a lack of agreement concerning the criteria that should be used to evaluate quality (Creswell, 2014). Several concepts such as crystallisation (Richardson, 2000b), transferability (Lincoln & Guba, 1985), and tacit knowledge (Altheide & Johnson, 1994) have been proposed to assess qualitative excellence. The conceptualisation of qualitative quality developed by Tracy (2010) was used to assess the quality of this work. According to Tracy, high qualitative research integrates the following eight markers of quality: "(a) worthy topic, (b) rich rigor, (c) sincerity, (d) credibility, (e) resonance, (f) significant contribution, (g) ethics, and (h) meaningful coherence" (2010, p. 837).

Worthy topic

According to Tracy, a good qualitative study should be "relevant, timely, significant, interesting, or evocative" (2010, p. 841). This research, which has the aim of gaining understanding of the ways in which sociocultural values and norms interact with existing and emergent material conditions t to shape "data journeys" (Bates et al., 2016) of patient data in

the UK health sector, is a relevant and significant contribution. First, it responded to the call of scholars in the Critical Data Studies field for conducting detailed empirical research to advance the understanding "of both the overall construction of data assemblages and their apparatus and individual elements" (Kitchin, 2014d, p. 6). In addition, this study is also significant in that it draws attention to the critical role that sociocultural values and norms play in shaping data movement. As pointed out in the literature review of this work, whereas a large number of studies exploring the circulation of data have been conducted, the majority of these studied have been conducted with a focus on management, infrastructural or technical issues (Bote & Termens, 2019; De Roo et al., 2016).

Rich rigor

Whilst quantitative studies are commonly valued for their exactitude, qualitative studies are characterised by their rich complexity of abundance (Winter, 2000). To generate richness it is necessary to integrate diverse contexts, data sources, and theoretical constructs (Weick, 2007). According to Tracy (2010), researchers that have developed a deep understanding of theories and collected enough data are better equipped to perceive complexity and nuance. A 'richly rigorous' (Tracy, 2010, p. 841) qualitative researcher is prepared to make appropriate decisions concerning contexts and samples that are useful to investigate specific matters. Throughout all the stages of this research project I conducted the review of literature to develop a good understanding of theoretical approaches and propositions that were useful for this research project.

Rigor, in addition to being connected to richness, also offers face validity, which pays attention to whether a research appears, on its face, to be appropriate and reasonable (Golafshani, 2003). In order to achieve rigor, I made decisions concerning data collection and analysis; only after a careful consideration about the potential implications of my decisions and following Tracy's advice, I pushes myself "beyond convenience, opportunism, and the easy way out" (2010, p. 841). I reflected whether the context and sample were appropriate to achieve the goals of the study.

During the data collection and analysis, I also reflected whether the data collected were enough to support significant claims (Tracy, 2010). Whereas there is no universal rule concerning how many qualitative interviews is enough because that depends on the particularities of the study, there is guidance to help deciding when to stop conducting interviews (Baker et al., 2012). For example, to determine if the interview data collected for this study was enough, I made sure to

stop the data collection only after exhausting all leads on every one of the data journeys and feeling confident that I captured sufficient depth on the phenomenon studied (Baker & Edwards, 2012).

A rigorous researcher should also adopt appropriate procedures during data collection and analysis. In this study, I was careful when selecting and applying data collection strategies, as well as when designing data collection instruments. In the process of designing the interview guides, I followed key principles to ensure that this instrument was adequate for my study. For example, I conducted desk research and literature review to gain a sound understanding about participants' context to be better prepared to define the questions to ask and understand what participants were trying to communicate (Roulston & Choi, 2018). In addition to this, I engaged in conversations with my supervisors to get feedback on the design process, and taking into account the recommendations provided by them, the instruments were refined and improved, until a final satisfactory interview guide was created.

Good practices when conducting interviews were adopted. For example, I practised the interviews before starting data collection, developed strategies to enhance rapport, and conducted all the interviews in a safe and secure environment. In order to achieve rigorous data analysis, Tracy (2010) recommends to explain the process followed to transform and organise data into the research report. In this study, I presented a detailed account of how I conducted the analysis in Section 3.7 of the Methodology chapter. I acknowledge that whereas rigour is a key element that must be present in high quality research, the presence of it does not guarantee a "perfect final product". However, rigor increases the odds for high quality (Tracy, 2010). In addition to this, the skills that are developed through rigorous practice provide qualitative researchers with better skills to apply in future research projects.

Sincerity

Sincerity in research can be accomplished through the combination of five key elements: honesty, self-reflexivity, transparency, vulnerability, and the auditing of data (Tracy, 2010). A researcher demonstrates sincerity by being honest and transparent about their biases, objectives, and shortcomings, and about the ways in which these played a role in the research project.

Self-reflexivity is an esteemed practice in qualitative research which is defined by being honest and authentic with "one's self, one's research, and one's audience" (Tracy, 2010, p. 842). Self-reflexivity motivates researchers to be open not only about their strengths but also about their weaknesses. In this study, I practiced self-reflexivity before embarking on data collection, but

also when negotiating access, conducting data collection and analysis, and reporting findings. I acknowledged that being a junior foreign researcher exploring data flows in the UK healthcare sector was a challenging endeavour for a number of reasons. First, I became familiar with the structure and functioning of the healthcare sector only one year before starting my research. Second, I did not have a network of contacts to assist me to get in touch with key informants, and English is not my first language. Recognising these weaknesses, I took actions from early stages of the research design to overcome these challenges. For example, I conducted desk research to familiarise myself with the structure, functions, and operation of the NHS. I also asked for recommendations from my supervisors and attended relevant training to design and improve my strategies to contact key informants and negotiate access. I also practiced my interviews in advance and made notes to myself, in order to be able to communicate what I wanted in the interviews. I conducted this study mindful of my shortcomings but convinced that I was capable to complete it in a successful way.

Transparency, which refers to being honest about the research process, was also demonstrated in this study. According to Tracy (2010) it is important that the researcher explain how they got into the context, how they immerse themselves in the field and their data collection practices. In section 3.6 of the Methodology chapter, I provided details about how interview participants were identified and approached, the interviews conducted (e.g. type of interview, structure and duration), and the interview transcription process. Another way to demonstrate transparency is by being open about the "challenges and unexpected twists and turns and revelation of the ways research foci transformed over time" (Tracy, 2010, p. 842). In this study, I was open about the fact that initially I intended to explore the journeys of patient data flowing to universities and pharmaceutical companies but that this was not possible due to challenges experienced in contacting key informants and negotiating access to pharmaceutical sites of data practice. In section 3.6.1.2 of the Methodology chapter, I explained the challenges experienced and explained how this impacted in the project.

Credibility

Credibility is concerned with the "trustworthiness, verisimilitude and plausibility of the research findings" (Tracy, 2010, p. 842). When a credible report is presented, a natural reaction from the readers is that they feel confident to take actions and make choices in line with it. In studies of a quantitative nature, credibility is achieved through accuracy, replicability, and consistency (Golafshani, 2003). Nevertheless, in qualitative research, these criteria cannot be

used in the same way to determine quality. Credibility in qualitative research is earned through practices such as triangulation or crystallisation (Tracy, 2010). Triangulation is a method used to demonstrate credibility, in which it is presumed that if several theories, data sources, or types of data gathered lead to the same resolution, findings are considered stronger and more credible (Denzin, 1978). Using triangulation can become problematic when one assumes that there is a singular truth that can be discovered. Researchers from interpretive paradigms have argued that the fact that all data point out to the same resolution does not mean that the insights presented are correct (Tracy, 2010). This is because findings of a research study are always influenced by the conditions of their generation, thus making direct comparisons of data collected through different means can be problematic (Bloor, 2001). Nevertheless, the practice of adopting multiple lenses and sources of data is important; an alternative way of doing this is the practice of crystallisation (Tracy, 2010). Crystallisation encourages a researcher to gather different types of data and employ various methods and more than one theoretical framework.

In this study I combined multiple methods to collect data. I conducted desk research and interviews and took reflective contextual field notes. In addition to this, I integrated two strands of thinking to conduct the data analysis and discussion of the findings, Data Journeys (Bates et al., 2016) and Data Valences (Fiore-Gartland & Neff, 2015). To further ensure credibility of this study, during all stages of this study, I held regular meetings with my supervisory team which served as peer debriefing (Lincoln & Guba, 1985). In these meetings, data collection and data analysis procedures were discussed and examined to ensure balance in data interpretation.

Resonance

The concept of resonance refers to the skills of research to reverberate and impact people in a meaningful way (Tracy, 2010). Written reports cannot offer direct insights into the anecdotal experiences of people (Schutz, 1967). However, the researcher can become involved in certain practices that are helpful to foster identification, empathy and reverberation of the research among readers who do not have a direct experience with the topic explored (Tracy, 2010, p. 844).

Resonance in research can be earned in different ways, such as evocative writing or formal generalisation and transferability. Whereas not all qualitative studies should achieve resonance through the same means, it is essential for “all high quality qualitative reports to have impact” (Tracy, 2010, p. 844). In this study, resonance was achieved through transferability and

naturalistic generalisation; these two processes are performed by the readers of the research report (Stake & Trumbull, 1982). It is important to consider that quantitative notions of generalisability are not useful and cannot be applied for qualitative research. The reason for this is that statistical generalisation needs random representational samples using data detached from specific contexts or situations (Tracy, 2010). Conversely, qualitative research engages in in-depth analyses that typically generate knowledge that is culturally and historically situated. Therefore, this knowledge cannot perfectly generalise to predict future practice.

Transferability is earned "when the reader experience the sensation that the story of the research overlaps with their own circumstances and in an intuitive way transfer the research to their own actions" (Tracy, 2010, p. 845). In order to invite transferability, I gathered and presented direct testimony from participants, provided rich descriptions and presented a report written in an accessible way.

Naturalistic generalisation is another path that leads to resonance. According to Stake and Trumbull (1982), the perception of personal knowing and experience results in enhanced practice. From this perspective, high quality research offers the reader a vicarious experience. In this study, I made an effort to present the report in a way that enabled the reader to make decisions drawing from their understanding of the scene.

Significant contribution

A high quality study should clearly point out the different ways in which it will "contribute to the understanding of social life" (Richardson, 2000a, p. 254). In the conclusion chapter of this study, I provided details of the contribution of this work. In terms of the theoretical contribution, I explained how the Data Journeys approach (Bates et al., 2016) in combination with the Data Valences (Fiore-Gartland & Neff, 2015) conceptualisation was applied to explore data journeys of patient data. Acknowledging that theoretical significance requires more than merely (re)applying existing theoretical frameworks (Tracy, 2010), I presented a reflection of how and why the combination of these two strands of thinking contributed to study the circulation of patient data. I also explained how building on Fiore-Gartland & Neff's work on data valences (2015) and the analysis of data collected developed and proposed an additional data valence.

This study is also heuristically significant in the sense that it motivates people to further explore additional aspects concerning the movement of data or the operation of socio-material factors in other settings (Tracy, 2010). In an attempt to increase heuristic significance, I provided

readers with a set of recommendations for future research. I also provided a set of recommendations for different actors such as university-based researchers and the public to engage in action. This aspect of heuristic significance overlaps to some extent with practical significance. With the objective of achieving practical significance, I captured the ways in which university-based researchers deal with situated issues and offers implications that could help participants in outlining a set of actions (K. Tracy, 1995).

Ethical

A number of practices that were discussed above such as transparency and self-reflexivity are key elements of ethical research. Nevertheless, ethics are not only a means, as they constitute a "universal end goal of qualitative research" (S. J. Tracy, 2010, p. 846). A variety of practices can be adopted to attend to ethics in studies of a qualitative nature; these include procedural, situational, relational and exiting ethics. Procedural ethics are those ethical actions established as compulsory by larger organisations or governing bodies (S. J. Tracy, 2010). In this study, procedural ethics dictated by The University of Sheffield Ethical Committee were followed. Details about this are provided in Section 3.9 of this chapter.

Situational ethics are concerned with ethical practices that originate from a careful reflection about the specific circumstances of a context (S. J. Tracy, 2010). In this study, apart from attending to situational ethics, I constantly reflected about my methods and the data worth revealing; I also sought advice from her supervisors to ensure that ethical decisions were made, taking into account the particular characteristics of this research project.

It is also important that the researcher take into account that ethical considerations are crucial when leaving the scene and sharing results, that is to say, exiting ethics. I agree that I cannot completely control how the work will read, be interpreted and used. However, I recognised that it was important to consider what is the best way to communicate the research in order to prevent unfair or unintended effects (Fine et al., 2000). For this reason, when writing the report, I was careful to not making misleading claims; I also received feedback from my supervisors to ensure the report was presented in a good way.

Meaningful coherence

Research that is meaningfully coherent links in a smooth way research design, and data collection and analysis strategies with theoretical considerations and situational ends (S. J. Tracy, 2010). To achieve meaningful coherence in this study, I used conceptualisations that

were in line with her research paradigm and research aims and objectives (S. J. Tracy, 2010). I also ensured that all the pieces of the research fitted together. This was evidenced in that the literature review situated the findings, the discussion of findings attended to the research question, and the conclusion interlinked in a meaningful way with the literature and data presented.

3.9 Ethical considerations

This section presents the ethical considerations that preceded the research design and implementation of this study. According to Bryman (2016), obtaining ethical approval from the university, the consent of participants and privacy protection and confidentiality are the most common ethical issues associated with contemporary social research. This section reflects on each of these aspects.

Ethical approval

As explained before, this research consisted of two phases. Ethical approval for Phase 1, which was exploratory in nature, was obtained on 25th April 2018. The approval for phase 2 (the main study) was obtained on 29th August 2018. In the two applications, relevant documents, including application form, participant information sheet and consent forms were submitted to be reviewed by the University of Sheffield's Research Ethics Committee. Details concerning research design, data collection and analysis strategies were clearly outlined in the application form. I confirmed that data collection would start only after obtaining approval from the university. Additional details about the ethics application and participant information sheets and consent forms can be found in the appendices 4-9.

Informed consent

Written consent was obtained to conduct all interviews (see appendices 5 and 7). Taking part was voluntary and participants were given the opportunity to withdraw from the study at any time without the need to provide any reasons. All participants were provided with relevant information about the aim and objectives of the research (see appendices 4 and 6). In addition to this, participants had the opportunity to ask questions about the project.

Confidentiality

All data collected in this research (e.g. audio recordings and transcripts) were properly handled to ensure confidentiality. All interviews were recorded and the researcher transcribed the audio into text for analysis. All digital data generated during the data collection of this research were

encrypted and stored on the Information School's research data drive, which was accessible only by me, my supervisors, the School's Examination Officer and ICT staff operating the facility. I also stored a password-protected and encrypted backup on my personal laptop. Data recorded in notebooks (e.g. reflective field notes) was held as securely as possible whilst in transit and stored in a locked office at other times. No highly sensitive information (e.g. names of participants) was recorded in notebooks.

Chapter 4. Description of data journeys examined

4.1 Introduction

This section presents a description of five selected data journeys of patient data that were explored as part of this study. Different material forms that 'journeys' of NHS patients' personal data take between different sites of practice, from their initial generation through to reuse for research purposes in different contexts, are presented. These data journeys are examples of the different ways in which patient data produced within the UK healthcare sector are reused for research in the academic context. As explained in the methodology chapter of this thesis, the researcher initially planned to follow the journeys of patient data flowing to two different types of sites: universities and pharmaceutical companies. However, significant recruitment difficulties were experienced with pharmaceutical companies. Due to the challenges experienced in recruiting key informants at pharmaceutical companies it was not possible to explore such journeys and a decision was made to focus only on reuse of patient data for research in universities.

These descriptive accounts are helpful to understand the series of activities and processes that are needed in order to obtain access to patient data for reuse in the academic context, and the roles of the different stakeholders that are involved at each stage of these data journeys. To be more specific, this chapter addresses the research sub-question number 2 of this study:

- What material forms do the 'journeys' of NHS patients' personal data take between different sites of practice, from their initial generation though to reuse for research purposes in different contexts?

Chapters 5, 6 and 7 will then go on to examine the socio-material drivers and frictions on these data flows. More specifically, they will address research sub-questions 3 and 4, which are:

- RQ3 In what ways do sociocultural values interact with existing and emergent material conditions to act as drivers to move data between different sites?
- RQ4 In what ways do sociocultural values interact with existing and emergent material conditions to act as frictions on efforts to move data between different sites?

The data journeys descriptions presented here were developed drawing initially from the accounts given by interviewees and informed with the analysis of additional sources provided by

the interviewees or publicly available such as data sharing agreements, annual reports, research protocols, and minutes of committees responsible for reviewing data access applications.

Each data journey includes a description of the path that patient data follows from the moment when it is initially produced to the moment when it is reused for research purposes. Since the focus of this study is to examine the interrelated socio-material factors that influence the movement of data, technical aspects are not discussed in detail. Each data journey description begins with a short summary of the research project in which patient data were used. This is followed by an explanation of the data production stage, which includes details on how patient data were initially generated. Later on, the arrangements that had to be put in place in order to allow the flow of data from the site of data production to sites of data reuse are explained. Finally, an explanation about how data are reused by researchers and practitioners is included.

The five data journeys described and visualised in this chapter were selected with the intention to capture a diversity of different types of data flows. Each of the five data journeys presented here centre their attention on different health issues, had different objectives and used data from different types of patients. The size of the research projects explored was also varied, projects of large, medium and small scale are included. Another difference is that each of the teams involved in these data journeys followed different processes to obtain access to the data they used, this means that they approached different data providers, therefore they engaged in negotiations with different stakeholders, and interacted with diverse intermediaries or organisations at national level, such as NHS Digital and Clinical Practices Research Datalink. It is important to mention that while additional data journeys were identified and partially followed, three of them were not integrated in this chapter because they were at very early stages. Therefore, it was not possible to produce meaningful descriptions and visualisations of them. As will be discussed in subsequent chapters (5 to 7), some minor differences in the socio-cultural and material factors shaping these data journeys were identified however, overwhelmingly the findings of this work show very similar patterns across all the journeys explored.

Producing the descriptions and visualisations of the data journeys explored was a step in the research design that allowed me to be confident and provide evidence that a variety of journeys were explored. They are also a significant contribution to knowledge because they bring to the surface five specific examples about the different ways in which patient data generated in the healthcare sector flow and are reused by external actors. These accessible visualisations and descriptions contribute to efforts of making data flows more transparent. This is particularly

useful given that flows of patient data produced in the healthcare sector are complex. As pointed out by others data flows in the healthcare sector are messy and difficult to understand (Harwich & Laycock, 2018) and while some documentation is publicly available that can illuminate some aspects of these data flows, this information tends to be difficult to navigate and overwhelming, and not very helpful to get a clear sense of who is accessing what and why.

4.2 Data journey 1: Using technology and data to improve the diagnosis and treatment of stroke

4.2.1 Research project summary

The research project *Using technology and data to improve the diagnosis and treatment of stroke*, hereinafter referred to as *Stroke Project*, was conducted by a team of researchers within the Connected Health Cities (CHC) Manchester programme (Connected Health Cities, n.d.-c). The CHC programme was funded by the Department of Health and it uses patient data to improve health of people who live in Northern England.

The *Stroke Project* aimed to develop a detailed overview about the journeys of patients between primary, secondary and community care in Manchester and Salford to gain a better understanding of the patient journey and to identify gaps in the care provided. Researchers involved in this project also had the objective of proposing improvements to support stroke patients and ensure the efficiency and adequate coordination of services.

This project consisted of three streams (Connected Health Cities, n.d.-c). For the purposes of this research project, only Stream 1 was explored. Stream 1 had the objective of identifying stroke mimics (false positive) and missed strokes (false negative). In order to achieve this objective, researchers from the Connected Health Cities programme used data generated at two different sites of data production: 1) historical ambulance data generated by the North West Ambulance Service (NWAS) and 2) Electronic Healthcare Record data from the Hyper Acute Stroke Unit (HASU) at Salford Royal NHS foundation Trust (SRFT).

4.2.2 Permissions required to access patient data

In the first place, a Data Sharing Agreement (DSA) between the Connected Health Cities project and the Salford Royal NHS Foundation Trust was set in place in order to establish a standard for the NWAS and HASU data sharing (Salford Royal NHS Foundation Trust, 2017). This DSA was signed with the objective of ensuring a secure and confidential sharing of

information between organisations. It specifies the purpose of sharing data from the SRFT with CHC; provides details about the conditions of use of the data to be shared; explains the measures that are required to be in place to ensure security of data; and establishes the retention period and format of the supplied data. Finally, this data sharing agreement also provides details about the legislation, guidance and information sharing principles with which all parties participating in the agreement are required to comply.

In order to obtain access to the data, the Connected Health Cities team made a Request of Data. This request included a summary of the project that CHC intended to carry out and specified what data they intended to access. This request also specified who the people accessing the data were, where the data were going to be held, the ways in which results of the study conducted were going to be disseminated, the period of time that data would be stored, and when they would be deleted.

4.2.3 Initial data production

Ambulance data: These data were produced by the North West Ambulance Service during the provision of medical services to a number of patients suspected of having a stroke in the North West of England. These patients received medical attention from NWAS paramedics and then were transported to the Hyper Acute Stroke Unit (HASU) at the Salford Royal NHS Foundation Trust. NWAS paramedics record data about a patient in a paper form called an *NWAS sheet.* Upon the patient arrival at the Hyper Acute Stroke Unit care at Salford Royal, the paramedics share a copy of the NWAS sheet with healthcare professionals (this sheet is in paper format because NWAS does not have electronic records).

Electronic Healthcare Record data: Patients admitted to the Hyper Acute Stroke Unit are assessed by healthcare professionals and receive medical care (Connected Health Cities, n.d.-c). Data generated about these patients from the time a patient is admitted until they are discharged are recorded in the Salford Royal NHS Foundation Trust Electronic Healthcare Record.

4.2.4 Technical arrangements prior to access data for secondary purposes

Data Linkage and removal of identifiable data: As previously explained, the North West Ambulance Service and the Hyper Acute Stroke Unit produce and manage data in different ways. The data generated by these two sites of data production are stored independently and in different formats. However, in order to undertake this project, it was necessary to combine data from these two different sources.

Staff at Salford Royal Foundation Trust scanned NWAS sheets into the Salford Royal Foundation Trust Electronic Patient Record and manually recorded structured data items into this database. The expenses were borne by the Connected Health Cities project.

In addition to this, analysts at Salford Royal Foundation Trust aggregated the postcodes according to a routine algorithm and removed all personal identifiable information from the records. The cost related to this process was also covered by Connected Health Cities.

Data transfer: NWAS and HASU care data were retrieved from the Electronic Patient Record and electronically transferred via File Transfer Protocol from the Salford Royal Foundation Trust server into the University of Manchester Trustworthy Research Environment (TRE) (Connected Health Cities, n.d.-c). This is a data analytics facility developed by Connected Health Cities Manchester and hosted by the University of Manchester campus, which is used as a secure place to store, analyse and process health data. The main objective of the TRE is to maintain confidentiality, integrity and availability of data. A number of security controls are used that prevent the misuse and unauthorised access of data. All data held by TRE is encrypted in transit and rest and it can only be accessed from virtualised workstations that do not have internet connection, because in this way the risk of data interception is minimised. The TRE virtual stations are equipped with data analysis tools such as R and STATA.

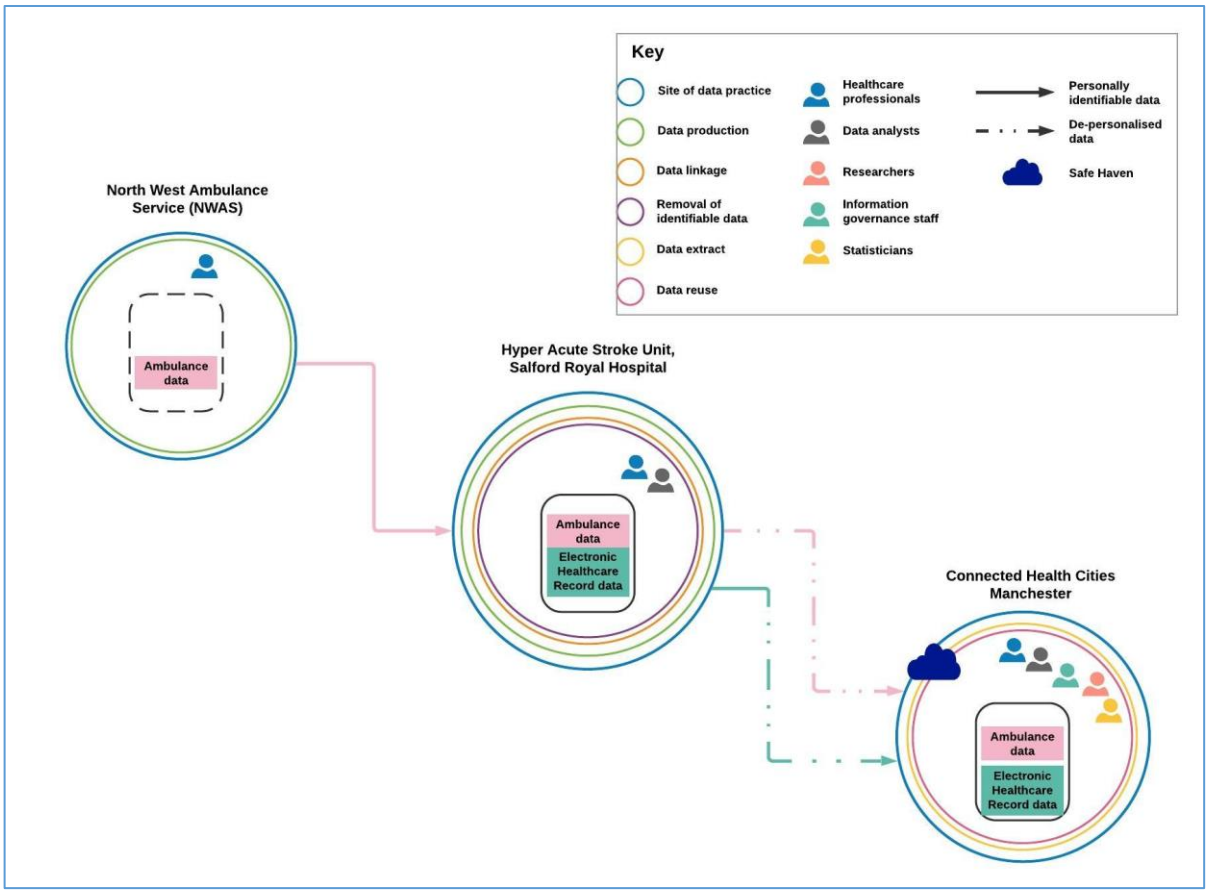
4.2.5 Data reuse

Once all the technical and information governance arrangements were in place, researchers from CHC accessed the data to conduct their intended research using the University of Manchester Trustworthy Research Environment. Staff from Connected Health Cities had access only to de-personalised data and agreed to handle data following Information Governance and Information Security rules in line with ISO27001, the NHS Information Governance Toolkit requirements, and the operating procedures of the Trustworthy Research Environment (Connected Health Cities, n.d.-b).

Table 3. Details of data generated per site of data production. Data journey 1

Details of data generated per site of data production	
Site of initial data production	**Details of data produced**
North West Ambulance Service	Demographic characteristics, episode and clinical information, details about injury, referred services and details about discharge (Salford Royal NHS Foundation Trust, 2017).
Hyper Acute Stroke Unit at Salford Royal Hospital	Patient demographics, diagnosis and relevant medical information, list of current medications (Salford Royal NHS Foundation Trust, 2017).

Figure 2. Data Journey 1: Using technology and data to improve the diagnosis and treatment of stroke



4.3 Data journey 2: Building Rapid Interventions to Reduce Antibiotic Prescription

4.3.1 Research project summary

The research project *Building Rapid Interventions to Reduce Antibiotic Resistance*, hereinafter referred to as *BRIT project*, was conducted with the aims of investigating why and when antibiotics are prescribed and studying how this impacts the health of patients (Connected Health Cities, n.d.-a). This research project consisted of a number of different stages which

involved the use of data from different sources. For the purposes of this project, only one stream was explored.

As part of the BRIT project, the team of researchers conducted a population-based cohort investigation. The cohort studied included patients of all ages who received an antibiotic prescription for incidental use (this refers to the prescription of antibiotics without any antibiotic prescribing in the past trimester) in the period from 2000 to 2016. For this stream of the research project researchers used GP data linked with other datasets (HES Admitted; ONS; Patient IMD and Practice IMD).

4.3.2 Permissions required to access patient data

Researchers from CHC obtained access to this data through Clinical Practice Research Datalink (CPRD). The Connected Health Cities team submitted an application to get access to anonymised patient level data held by CPRD using a Protocol Application Form provided and designed by CPRD. The protocol submitted was reviewed by the Independent Scientific Advisory Committee (ISAC) (Clinical Practice Research Database [CPRD], n.d.-c).

4.3.3 Initial data production

General Practice data: The GP data used for this project come from GP practices using the VISION or EMIS software systems that have accepted to share their patients' data with CPRD. Patient data fully coded and anonymised is recorded into the Electronic Health Record of GP practices and flows to CPRD on a regular basis. GPs record patient data directly into the EHR of each patient that receives healthcare at their practice. Before transferring data to CPRD, the GPs' software suppliers that act as data processors (EMIS or VISION) remove personal identifiers e.g. name, complete DOB, postcode and NHS number. Personal data of patients who have opted-out at sharing their patient data are not transferred to CPRD (Connected Health Cities, n.d.-a).

HES Admitted Patient Care Data (HES APC): All NHS providers (acute hospital trusts, primary care settings and mental health trusts) generate data about all admissions to or attendances at English NHS healthcare providers. HES APC data brings together the whole set of hospital episode information "(date of admission, date of discharge, diagnoses, specialists seen and medical procedures) for each associated patient with an in-patient record" (Clinical Practice Research Database [CPRD], n.d.-a, sec. HES Admitted Patient Care data). NHS

healthcare providers routinely share these data with NHS Digital (Clinical Practice Research Database [CPRD], n.d.-a).

Death registration data (ONS): Data about all deaths "registered in England and Wales is collected by the Local Registration Service and the General Register Office (GRO)" (Clinical Practice Research Database [CPRD], n.d.-b, p. 4). Death data is uploaded by registrars on the Registration Online database. The majority of the data is usually provided by the person reporting the death, whereas the cause of death is commonly provided by the Medical Certificate of Cause of Death (MCCD) completed by a healthcare professional once the death is certified (Clinical Practice Research Database [CPRD], n.d.-b). The Office for National Statistics provides mortality data to NHS Digital on a monthly basis (NHS Digital, n.d.-d).

Patient and Practice Index of Multiple Deprivation (IMD): This is an integrated measure derived from seven indicators covering a number of factors of material deprivation: "earnings, employment, educational background and skills development, health deprivation and disability, barriers to housing and services, crime, and living environment" (Ministry of Housing, Communities and Local Government, 2019, p. 2). The Ministry of Housing, Communities and Local Government is responsible for calculating the IMD in England (Ministry of Housing, Communities and Local Government, 2019).

4.3.4 Technical arrangements prior to access data for secondary purposes

Data linkage and removal of identifiable data: As mentioned previously, for conducting the BRIT project**,** researchers required the GP data to be linked with other datasets (HES Admitted; ONS; Patient IMD; and Practice IMD). The data linkage for this project was conducted by NHS Digital on behalf of CPRD, which is the statuary body in England with legal authorisation to collect personal identifiable data (Connected Health Cities, n.d.-a).

Software providers of electronic health records share patient data with Clinical Practice Research Datalink and NHS Digital. CPRD receives de-identified data, whereas NHS Digital receives CPRD pseudonyms, NHS numbers, DOB, and postcode. Data custodians of external datasets submit to NHS Digital personally identifiable information (NHS number, gender, DOB and postcode) and a depersonalised patient record identification code. After this, NHS Digital links the datasets by using identifiers (Padmanabhan et al., 2019).

Once NHS Digital receives CPRD and external datasets, it generates a 'linkage file' by matching patient identifiers from two datasets (this linkage file does not include any direct

patient identifier). After this, NHS digital shares the linkage file with CPRD; this permits CPRD to conduct the linkage of datasets without requiring any personal identifiers to merge the data. Following this, NHS digital sends to CPRD a de-identified linked cohort file. Finally, CPRD shares anonymised linked data with researchers.

4.3.5 Data reuse

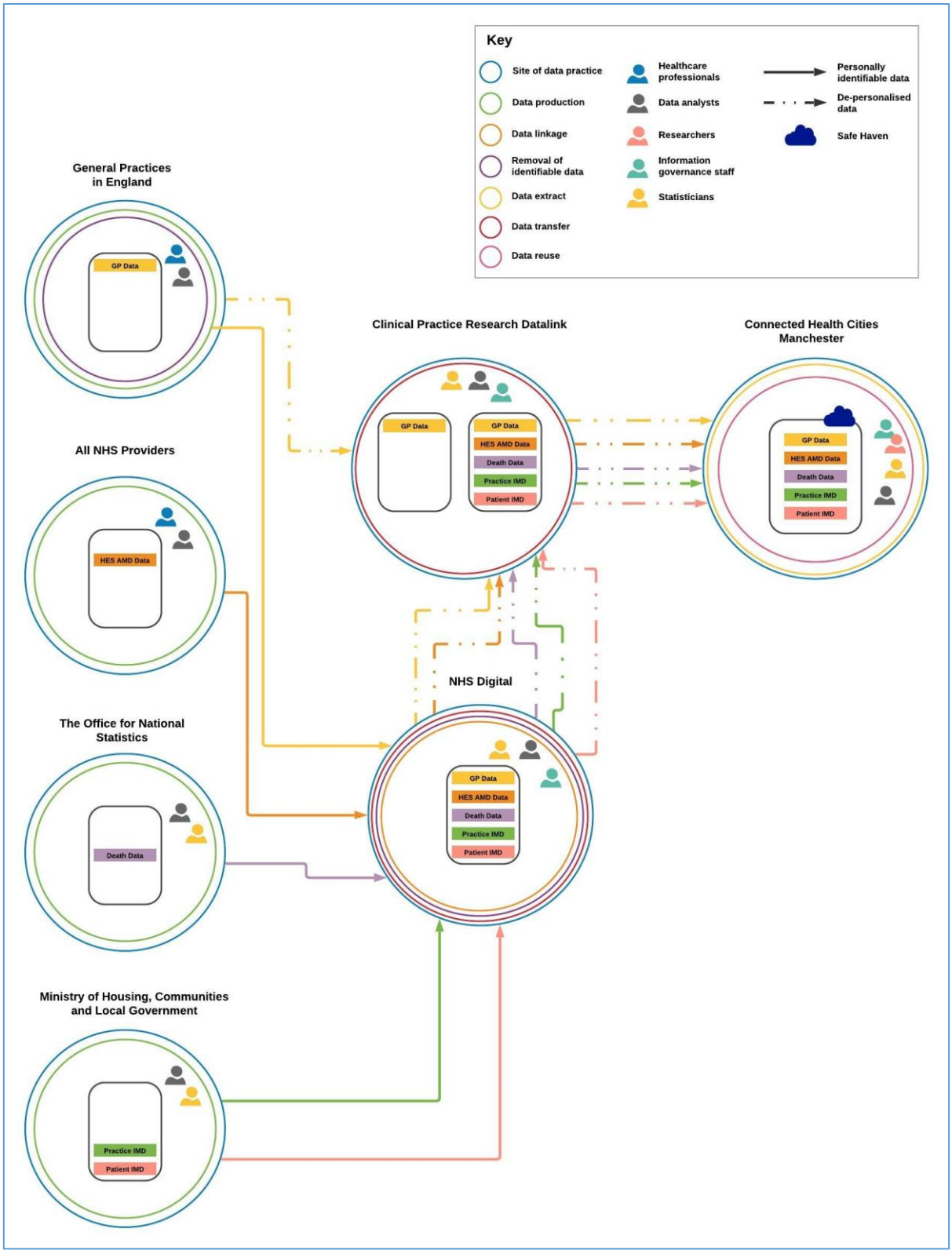
CPRD provided anonymised linked data to researchers at the University of Manchester. Researchers used an analytical strategy to organise in clusters diagnostic data connected to prescription of antibiotics. The outcome of this analysis allowed researchers to identify the factors connected to the levels of antibiotic prescribing between different healthcare settings. In addition to this, researchers used these data to compare the incidence of infections between antibiotic users and non-users after one month of being admitted to the hospital (Connected Health Cities, n.d.-a).

In order to promote a better understanding of the factors that affect antibiotic prescribing, the team of researchers of the BRIT project created two dashboards: a National Antibiotic Prescribing Tool and the GP Antibiotic Prescribing Dashboard. The National Antibiotic Prescribing Tool can be accessed by a number of stakeholders, such as policymakers, to gain a better understanding of the factors that influence the antibiotic prescribing profile of the UK. The GP dashboard is used by GPs and healthcare professionals, as a tool to compare their own patterns of antibiotic prescription at national and regional level. The GP dashboard enables GPs to make comparisons of their own antibiotic prescribing patterns with regional and national figures.

Table 4. Details of data generated per site of data production. Data journey 2

Details of data generated per site of data production	
Site of initial data production	**Details of data produced**
General Practices	Demographic details, symptoms developed and diagnoses, vaccines administered, lab results, referrals to hospitals.
NHS providers (acute hospital trusts, primary care trusts, and mental health trusts)	Admission and discharge dates, diagnoses, specialists seen under and procedures undertaken.
General Register Office	Date of registration of death, date of death, underlying cause of death.

Figure 3. Data journey 2: Building rapid interventions to reduce antibiotic prescription



4.4 Data journey 3: Long-term outcomes of urinary tract infection in childhood

4.4.1 Research project summary

The *Long-term outcomes of urinary tract infection (UTI) in childhood (LUCI)* study was conducted by a group of researchers at the Centre for Trials Research at Cardiff University (Cardiff University, n.d.). This study had the objective of investigating long-term outcomes of children who have experienced a urinary tract infection at any point during their first five years of life. This study started in October 2016 and ended in June 2019. The LUCI study aimed to answer whether children who have suffered a urinary tract infection aged have poorer outcomes compared with children who have not experienced a UTI. The NHS Wales Research Ethics Committee and the Health Research Authority's Confidentiality Advisory Group approved this research protocol (Lugg-Widger et al., 2019).

This project involved the utilisation of two datasets of children. Dataset 1 consisted of data from children living in Wales "(hospital, general practice, and microbiology)" (Lugg-Widger et al., 2019, p. 1) Dataset two included data about children who participated in two observational studies, Diagnosis of Urinary Tract infection in Young children (DUTY) and Epidemiology of Urinary Tract Infection (UTI) in Children with Acute Illness in Primary Care (EURICA) using urine samples for children with acute illness receiving medical attention at primary care settings (Cardiff University, n.d.). The EURICA study recruited participants from GP practices in Wales. DUTY recruited participants from practices in England and Wales.

4.4.2 Permissions required to access patient data

All data used for this research project was accessed through the Secure Anonymised Information Linkage databank (SAIL) (SAIL Databank, n.d.-c). SAIL is a national data safe haven for anonymised health and administrative data about residents of Wales (SAIL Databank, n.d.-c). This research was reviewed and approved by the Information Governance Review Panel (IGRP); this panel is composed of representatives of regulatory and professional organisations and members of the public (SAIL Databank, n.d.-a).

In addition to this, it was necessary to obtain approval from the Health Research Authority's Confidential Advisory Group (CAG) in order to be able to use data from children who participated in DUTY and EURICA for the LUCI project (Lugg-Widger et al., 2019).

The Independent Group Advising on the Release of Data (IGARD) granted approval to access EURICA and DUTY data. All children guardians' and parents' had the opportunity to withdraw from the study. They were given the opportunity to contact the research team via email, text or phone to communicate/inform of their dissent. Participants who registered their dissent were removed from the datasets before starting the LUCI study.

4.4.3 Initial data production

Dataset one. Data from children living in Wales. Data of children corresponding to dataset 1 come from the following different sources: general practices, "Patient Episode Database for Wales (PEDW), Welsh Demographic Service (WDS), Welsh Electronic Cohort of Children (WECC)" (Lugg-Widger et al., 2019, p. 3), outpatient data, and a data repository that holds data produced by the Laboratory Information Management Systems in Wales (InterSystems Corporation, n.d., para. 1).

General practice data: Clinicians at general practices across Wales record patient data into the electronic health record of each patient. The patient data recorded at GP practices include: diagnoses, symptoms presented, results of tests, treatments, and specialist referrals (SAIL Databank, n.d.-c). These data are regularly submitted to SAIL Databank in an anonymised version.

Patient Episode Database for Wales (PEDW): Healthcare professionals at hospitals across Wales collect administrative data e.g. speciality of care, dates of admission, and dates of discharge from the central Patient Administrative System. Once children are discharged, all the patient record notes are translated into healthcare terminology. NHS providers submit this data to NHS Wales Informatics Service (NWIS), which is in charge of managing this dataset. These data are regularly provided to SAIL Databank by NWIS in an anonymised version (SAIL Databank, n.d.-d).

Welsh Demographic Service (WDS) Data: Administrative data about children living in Wales that are users of the NHS medical services are retrieved from general practices by NWIS via Exeter System. This data is regularly provided to SAIL Databank by NWIS in an anonymised version (SAIL Databank, n.d.-e).

Outpatient Data (OPD): Staff at hospitals across Wales collect administrative data about hospital outpatient appointments using the central Patient Administrative System. Data gathered include speciality of care, appointment data and record of attendance. NHS providers

submit this data to NHS Wales Informatics Service, which is in charge of managing this dataset. These data are regularly provided to SAIL Databank by NWIS in an anonymised version (SAIL Databank, n.d.-b).

Microbiology data: Urine microbiology test results produced at laboratories in Wales are extracted using a database tool developed by PHE (Public Health England, 2019).

Dataset two. Data from children born and resident in England. Dataset two included data from participants in the DUTY and EURICA studies. Bristol University sponsored DUTY and Cardiff University sponsored EURICA. The sponsors of both studies collected data of ill children (aged 5 years old or younger) diagnosed with a UTI infection in primary care (Cardiff University, n.d.).

Clinical and Demographic Data: This type of data was gather by the researchers working at DUTY and EURICA projects. In addition to this, the sponsors of both studies requested samples of urine from all children who participated in the LUCI to be analysed in NHS laboratories.

Hospital Episode Statistics (HES): Inpatient and outpatient data about children who took part in the EURICA and DUTY were initially collected by NHS service providers. For the purposes of this study, HES data were requested to NHS Digital. These data included diagnoses, medical procedures and duration of episodes (NHS Digital, 2019a).

4.4.4 Technical arrangements prior to access data for secondary purposes

Dataset one. Data from dataset one (routinely collected data from children living in Wales) were combined using SAIL. Only data providers are able to see identifiable datasets; therefore SAIL cannot receive identifiable records. Thus in order to conduct the data transfer, the data provider divided the dataset into two parts: demographic details and content (SAIL Databank, n.d.-a). Demographic details, including name, address, and DOB were shared with NHS Wales Informatics Service (NWIS). The content data (medication, diagnoses) were sent directly to SAIL databank. At NWIS, NHS numbers and identifiers were removed and replaced with an anonymised linking field (ALF). The ALFs, in conjunction with gender and minimal aggregated data, were then sent to the SAIL databank to be recombined with the "content" elements of the datasets.

Dataset two. Cardiff University transferred identifiers of participants from dataset 2 (EURICA and DUTY) to NHS Digital and NWIS for data matching (Lugg-Widger et al., 2019). Data

from EURICA was matched by NWIS, whereas data from DUTY was matched by both NWIS (children from Wales) and NHS Digital (children born in England. After conducting the matching, de-identified data from both studies were transferred to SAIL. The clinical data from participants from both studies were directly transferred to SAIL. The data received by SAIL included clinical data, de-identified demographic variables, and the ALF.

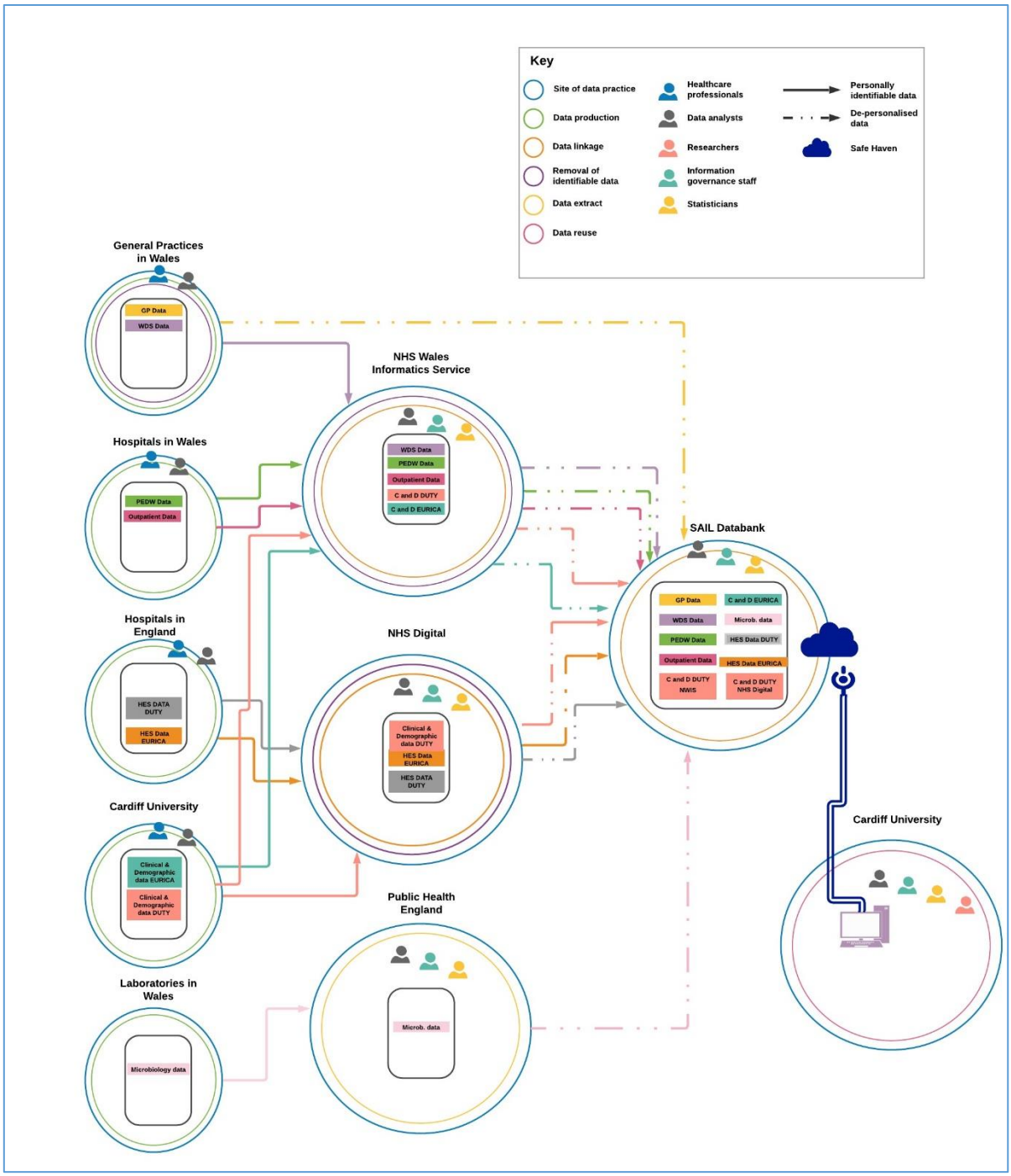
4.4.5 Data reuse

Researchers accessed patient data via the SAIL Databank Gateway platform. Thanks to the use of remote access technology, researchers were able to access data from their own desktop. Researchers were allowed to view data but not to extract it from SAIL Databank. Before accessing data, researchers had a training session for safe conduct with data (Safe-Researcher) and agreed to comply with the SAIL data access agreement. The outputs of this research project included a protocol paper and an additional paper answering the research questions of the study. The findings of this study are expected to be useful for healthcare professionals and policymakers and are expected to influence the treatment of UTI in children.

Table 5. Details of data generated per site of data production. Data journey 3

Details of data generated per site of data production	
Site of initial data production	**Details of data produced**
General Practices in Wales	Diagnoses, signs, symptoms, test results, prescribed treatment, referrals to specialists. Administrative data (address, registration history).
Hospitals in England and Wales	Admission and discharge dates, diagnoses, and operations performed.
Cardiff University	Clinical and demographic data of ill children from Wales and England.
Laboratories in Wales	Microbiology culture results from all laboratories.

Figure 4. Data journey 3: Long-term outcomes of urinary tract infection in childhood



4.5 Data journey 4: Linking electronic health records with smartphone data to predict outcomes in psychotic disorders

4.5.1 Research project summary

The project *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders* is a project that is currently in process, conducted by a team of researchers at the Maudsley Biomedical Research Centre, King's College London (UK Research and Innovation [UKRI], n.d.). This research centre was created through an initiative between the "South London and Maudsley (SLaM) NHS Foundation Trust and the Institute of Psychiatry, Psychology and Neuroscience at King's College London" (Maudsley Biomedical Research Centre, n.d.-a, para. 1). The data used for this project was generated in the SLaM NHS Foundation Trust, which offers mental health care and treatment to 1.2 million people who live in south London (Lambeth, Southwark, Lewisham and Croydon boroughs).

This project investigates clinical outcomes in mental disorders and it includes two stages. For the first stage of the project (in which researchers are working at the present time) researchers are analysing electronic health record data of patients with a first episode of psychosis to investigate the possibility of using these data to predict relapses at individual patient level (UK Research and Innovation [UKRI], n.d.).

The data that are being used for this stage of the research project come from psychiatric hospital admission data. The second stage of this project has not started yet and researchers are in the process of obtaining ethical approval. This second stage involves the utilisation of smartphone data. For this stage, the researchers will investigate the possibility of combining smartphone data with electronic health record data to predict relapses.

4.5.2 Permissions required to access patient data

A particularity of this project is that no intermediaries are involved to obtain access to the patient data used for this project. To be more specific, it was not necessary to apply for National Ethical approval or obtain permission to access patient data from external organisations such as NHS Digital or CPRD.

However, in order to access patient data for this project researchers were required to seek the approval of the Clinical Record Interactive Search (CRIS) Oversight Committee, which is a local committee that ensures that all applications for research projects are in line with ethical

and legal requirements (Maudsley Biomedical Research Centre, n.d.-b). This committee is led by a service user and has a representation from the SLaM Caldicott Guardian.

4.5.3 Initial data production

South London and Maudsley NHS Foundation Trust data: Clinicians at the South London and Maudsley NHS Foundation Trust record patient data into the electronic health record (EHR). Data recorded by clinicians in the EHR then flows to the SLaM Biomedical Research Centre (BRC) case register, which is a dynamic dataset based on the EHR of this Trust (NHS Health Research Authority, n.d.). The case register is updated every 24 hours from the live EHR and contains records going back to 2006. Clinical data, including free text, from SLAM are available to conduct research in the form of anonymised datasets.

4.5.4 Technical arrangements prior to access data for secondary purposes

Removal of identifiable data: Once the patient data is recorded in the EHR it runs through various pipelines to structure it, and to remove the identifying material from it. Data is then stored in the Clinical Record Interactive Search System (CRIS), which is a digital system (SQL database) used within the NIHR Maudsley Biomedical Research Centre (BRC) that enables researchers to access anonymised patient data extracted from the Electronic Health Record System of the this Trust (Maudsley Biomedical Research Centre, n.d.-b). CRIS removes patient data such as names and addresses from health records to prevent the identification of patients. This SQL database can be interrogated by researchers in two ways: 1) using a web-based application, CRIS FAST or 2) using SQL Microsoft Structured Query Language.

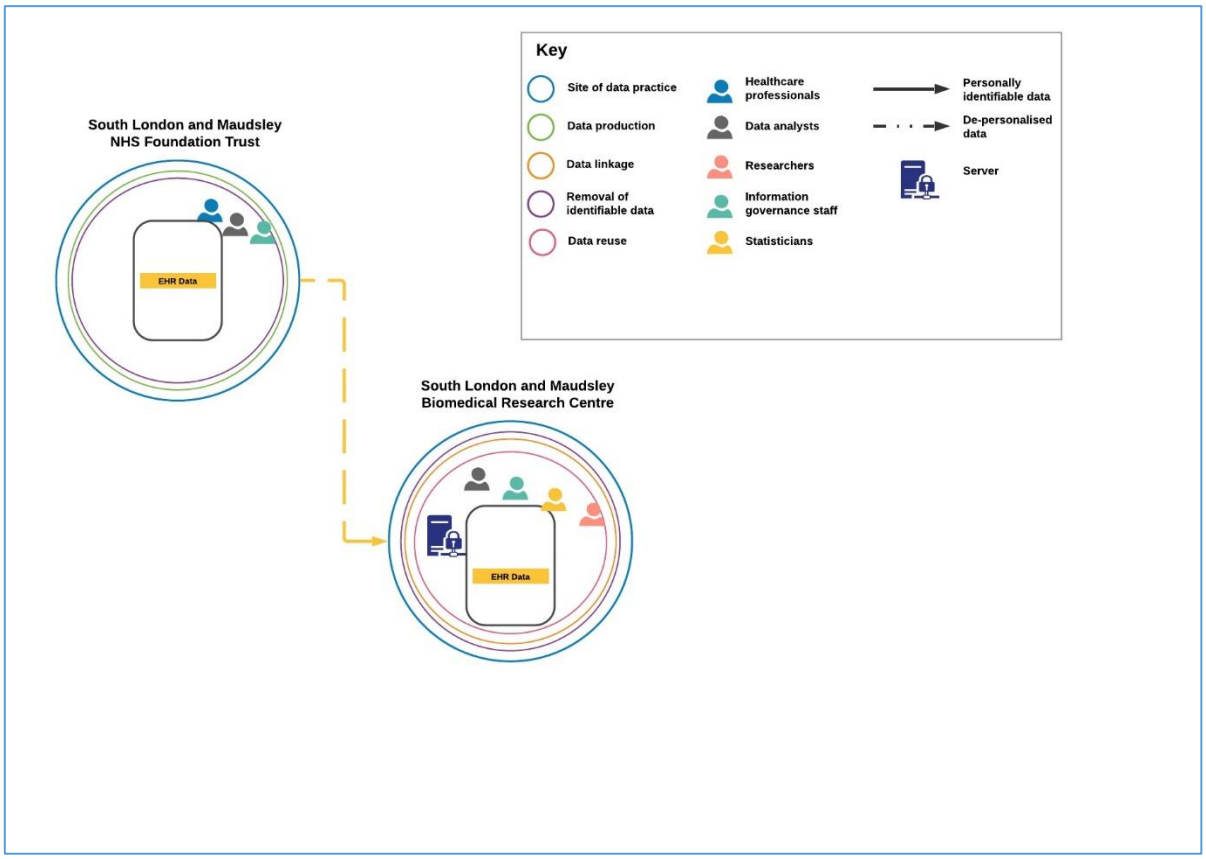
4.5.5 Data reuse

All the data analysis for this research project was conducted within the Trust firewall. Researchers had access to notes and other data in text format. They accessed these data through CRIS SQL using Natural Language Processing (NLP). For the first part of the study, researchers applied natural language processing software to the free text notes to put out clinical information (e.g. symptoms presented by the patients, diagnoses, medications). These data will be used to develop a prediction model to predict relapses.

Table 6. Details of data generated per site of data production. Data journey 4

Details of data generated per site of data production	
Site of initial data production	**Details of data produced**
South London and Maudsley Biomedical NHS Foundation Trust	Psychiatric hospital admission data, clinical data (including free text).

Figure 5. Data journey 4: Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders



4.6 Data journey 5: Age inequalities and inequities in the cancer pathway

4.6.1 Research project summary

Bowel Cancer Intelligence (BCI) UK is a research collaboration hosted by the University of Leeds and funded by Cancer Research UK (University of Leeds, n.d.). BCI is creating CORECT-R, which is a data repository that contains a broad variety of cancer e.g. diagnosis, treatment and outcome (University of Leeds, n.d.). CORECT-R gathers National Cancer Registration and Analysis (NCRAS) data to a number of other datasets collected at a national level, such as Hospital Episode Statistics (HES), Diagnostic Imaging Dataset, and National Cancer Waiting Times Monitoring Data (UK Colorectal Cancer Intelligence Hub, n.d.-b). The data stored in CORECT-R will be available to researchers who comply with the relevant ethical and regulatory

frameworks. At the time when this data journey was explored the data on CORECT-R was only available to researchers based at the Leeds Institute for Data Analytics (LIDA).

Researchers at the CORECT-R programme are currently working on different research projects in three key research areas of bowel cancer: supporting earlier diagnosis, tackling inequalities, and optimising cancer research. For the purposes of this work, only one data journey was explored in detail. The journey explored corresponds to a project in the area of "Tackling inequalities" called *Understanding age inequalities and inequities in the cancer pathway*. The project explored had the main objective of studying the use of radical rectal cancer treatments, and their related outcomes across England (Birch et al., 2019).

The data used for this project included data generated within and outside the UK healthcare sector. Data generated within the healthcare sector corresponded to data of individuals diagnosed with a first primary rectal cancer in England between April 2009 and December 2014 (Birch et al., 2019). Data generated outside the healthcare sector comes from the Quality of Life of Colorectal Cancer Survivors in England: Patient Reported Outcome Measures Survey (PROMS). More details about these data are provided in the following section (National Health Service [NHS UK], n.d.).

4.6.2 Permissions required to access patient data

The team of researchers at LIDA obtained approval in June 2018 from the Research Ethics Service of the Health Research Authority to develop CORECT-R.

In addition to this, the Public Health England Office for Data Release (ODR) and the University of Leeds signed a data sharing agreement, which enables the use of de-personalised data for bowel cancer research. The ODR provides guidance for evaluating data requests to use data produced by Public Health England for secondary purposes (Public Health England, n.d.). Data access requests submitted to the ODR are subject to confidentiality conditions in agreement with key regulatory and information governance frameworks e.g. the General Data Protection Regulation, Caldicott principles and, and the national data opt-out programme (Public Health England, n.d.).

Researchers working on this project submitted an application to the ODR, which assessed the application and granted approval in June 2017. It was agreed that identifiable data would be stored in a data safe haven provided by Public Health England. Only de-personalised data would be extracted from this secure environment to be stored in the University of Leeds, where

it could be reused by research groups from across the UK (UK Colorectal Cancer Intelligence Hub, n.d.-a).

4.6.3 Initial data production

This project used data from cancer patients generated by a number of NHS service providers, which included Acute NHS Trusts, Acute NHS Foundation Trusts, NHS Primary Care Trusts, NHS Care Trusts, NHS Treatment Centres, Private and Independent Care Providers (with ODS codes and NHS.net connectivity), and Private and Independent Screening Service Providers (with ODS codes and NHS.net connectivity) (Ambrose, 2010). These data are routinely collected, organised in different datasets and stored by the National Cancer Registration and Analysis Service (the systematic collection of these data is known as cancer registration) (Public Health England, 2020).

In addition to this, the project used data from the Quality of Life of Colorectal Cancer Survivors in England: Patient Reported Outcome Measures Survey (PROMS) (National Health Service [NHS UK], n.d.), which is held by NCRAS but generated outside the healthcare sector. A brief explanation of how these data are initially generated and how they flow to NCRAS is presented as follows:

Cancer registration: NHS providers, such as NHS Trusts, collect data about patients who have been diagnosed with cancer in England, including data on the patient, their diagnosis, tumour characteristics and details of the care and treatment received. These data are submitted to NCRAS on a monthly basis and recorded into the Cancer Registration dataset (Public Health England, 2020). Public Health England acts as the data controller of this dataset.

Cancer waiting times data: All providers (Acute Trusts, Care Trusts and independent providers) offering cancer care and treatment (NHS Digital, 2017) submit data to NCRAS on referrals, first outpatient appointments, diagnosis and treatments derived from patient care activity (NHS Digital, 2018b). NHS England acts as the data controller of this dataset.

Radiotherapy dataset: All NHS Acute Trusts providers of radiotherapy services in England submit to NCRAS data about radiotherapy (teletherapy and brachytherapy given using automated remote afterloading machines) delivered to patients from 1st April 2009 (NHS Digital, n.d.-e). Data about all other brachytherapy for the treatment of malignant disease provided in England to patients in NHS settings, or in external settings where patient care is

paid by the NHS, are also submitted to NCRAS. These data are then recorded into the radiotherapy dataset. Public Health England acts as the data controller of this dataset.

Hospital Episode Statistics (HES): NHS hospitals collect data when patients receive treatment and care, including details of admissions Accident and Emergency attendances and outpatient appointments (NHS Digital, 2019a). These data are shared with NHS Digital to be processed and then are sent back to healthcare settings as the Secondary Uses Service (SUS) (NHS Digital, 2018c). These same data are processed by NHS Digital for uses beyond the direct care of patients (secondary uses) and recorded into the Hospital Episode Statistics (HES) dataset. HES data include details about diagnoses and procedures, demographics, administrative details (e.g. dates of admission and dates of discharge), and geographical details (e.g. patient's address).These data flows to NCRAS and NHS Digital acts as the data controller of this dataset.

Cancer screening programmes dataset: is integrated by anonymised data retrieved from National Health Application and Infrastructure Services (NHAIS) systems. The reports that integrate the Cancer Screening Programme National Statistics provide high level statistical summary data on key aspects of the national cancer screening programmes; however, these reports do not include any individual patient or GP data (NHS Digital, 2018a).

Route to diagnosis dataset: This dataset is integrated by derived fields. "Administrative HES data are combined with Cancer Registration Data, Cancer Screening Programmes, and Cancer Waiting Times" (National Cancer Registration and Analysis Service, n.d., para. 2). Using data derived from these datasets, all cancer cases recorded in England diagnosed from 2006 to 2013 are classified into one of the existing 'Routes to Diagnosis' and recorded into the Route to Diagnosis dataset (National Cancer Registration and Analysis Service, n.d.). Public Health England acts as the data controller of this dataset.

Quality of Life of Colorectal Cancer Survivors in England: Patient Reported Outcome Measures Survey (PROMS): This dataset does not include data collected at NHS service provider settings; it only includes responses of a questionnaire sent to individuals alive between twelve and thirty-six months after being diagnosed with colorectal cancer. Details about these individuals were retrieved from the National Cancer Registration Service (Public Health England, 2017). Data in this dataset include: type and length of treatment received by the patient; patient's perceptions concerning the physical and mental effect of the cancer; details

about the care provided to the patient in primary and secondary care settings; and demographics (Glaser et al., 2015).

4.6.4 Technical arrangements prior to access data for secondary purposes

Identifiable data are stored in a data safe haven within Public Health England with restricted access (UK Colorectal Cancer Intelligence Hub, 2021). The aforementioned datasets are then linked by a member of the Bowel Cancer Intelligence Hub UK team. The linkage is conducted within the Cancer Analysis System (CAS) by a member of the Bowel Cancer Intelligence Hub UK team (the data manager of the project). De-identified patient data is extracted from this system to be stored in the University of Leeds, where it can be reused by UK-based research groups (NHS Health Research Authority, 2018).

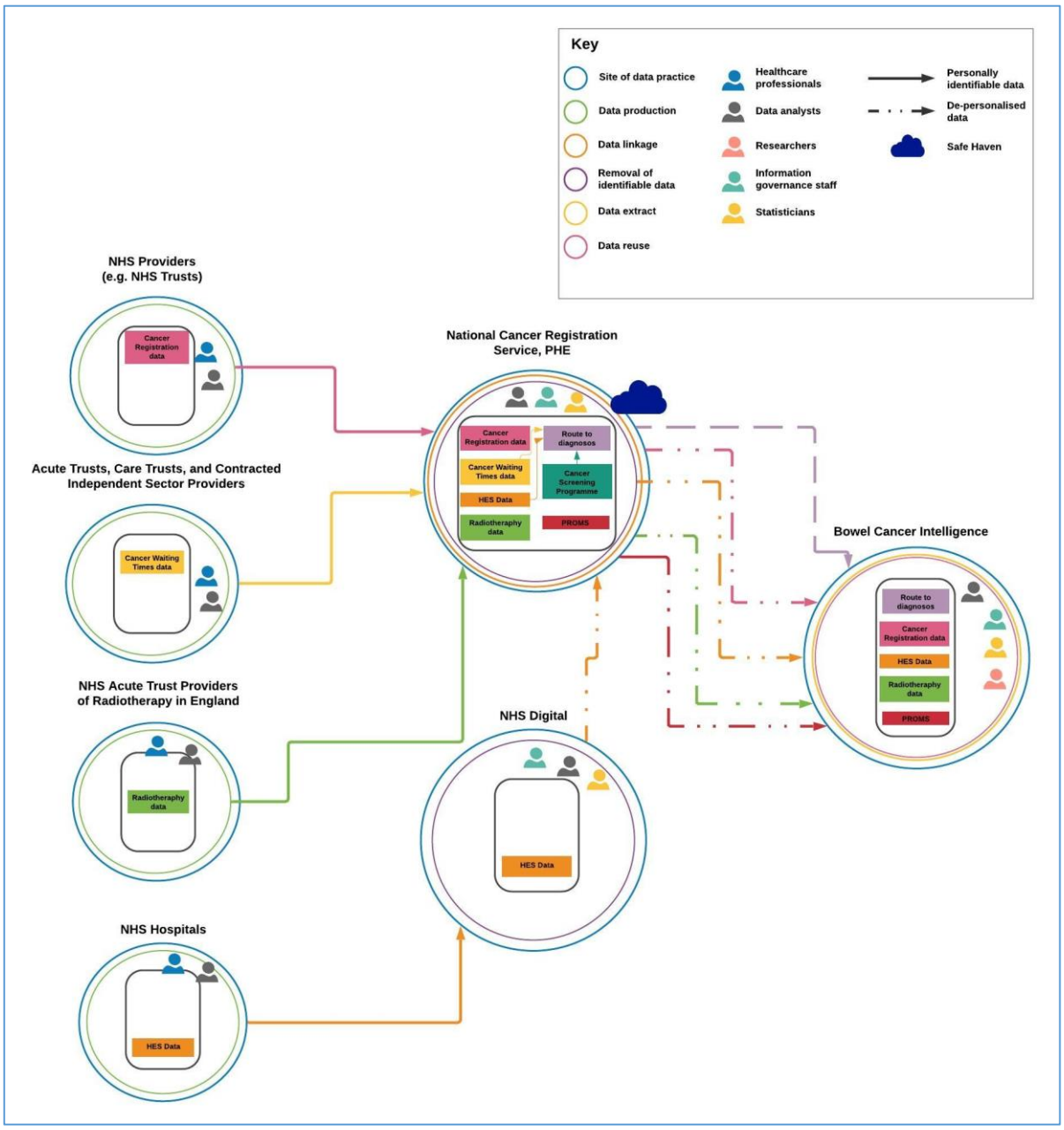
4.6.5 Data reuse

Researchers had access only to de-identified patient data. As part of this project, they conducted statistical analyses (e.g. use of multilevel binary logistic regression, use of Spiegelhalter method, adjusted logistic regression models). The statistical analysis of this project was conducted using STATA 15.0.

Table 7. Details of data generated per site of data production. Data journey 5

Details of data generated per site of data production	
Site of initial data production	**Details of data produced**
All NHS Providers	Data about patients who have been diagnosed with cancer in England. Diagnosis, tumour characteristics and details of the care and treatment received.
NHS providers (acute hospital trusts, primary care trusts, and mental health trusts)	Referrals, first outpatient appointments, diagnosis and treatments derived from patient care activity.
NHS Acute Trust Providers of Radiotherapy in England	Data about teletherapy and brachytherapy given using automated remote afterloading machines delivered to patients from 1st April 2009.
NHS Hospitals	Details about diagnoses and procedures, age group, ethnicity, administrative details, and geographical information.

Figure 6. Data Journey 5: Understanding age inequalities and inequities in the cancer pathway



4.7 Chapter overview

This chapter presented detailed descriptions and diagrammatic representations of five selected data journeys of patient data that were followed as part of this study. These descriptions were developed drawing from the accounts given by interviewees and document analysis of additional sources provided by the interviewees or publicly available. Through these descriptions it was possible to understand what material forms the 'journeys' of NHS patients' personal data take between different sites of practice, from their initial generation though to reuse for research purposes in different contexts. The journeys followed are summarised below.

Data journey 1, *Using technology and data to improve the diagnosis and treatment of stroke* was conducted as part of the Connected Health Cities (CHC) Manchester programme. This project had the objective of proposing improvements to support stroke patients. Datasets reused in this project were produced at two different sites of data practice, the North West Ambulance Service and the Hyper Acute Stroke Unit at Salford Royal Hospital.

Data journey 2, *Building Rapid Interventions to Reduce Antibiotic Prescription* was conducted as part of the Connected Health Cities (CHC) Manchester programme. This project was conducted with the aims of investigating the factors that lead to the prescription of antibiotics and its related health outcomes. Datasets reused in this project were produced at three different sites of data practice: general practices, NHS providers (acute hospital trusts, primary care trusts, and mental health trusts), and the General Register Office.

Data journey 3, *Long-term outcomes of urinary tract infection in childhood (LUCI)* was conducted by a group of researchers at the Centre for Trials Research at Cardiff University. This study had the objective of investigating long-term outcomes of children who have experienced a urinary tract infection during their first five years of life. Datasets reused in this project were produced at four different sites of data practice: general practices in Wales, hospitals in England and Wales, Cardiff University, and laboratories in Wales.

Data journey 4, *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders* was conducted by a team of researchers at the Maudsley Biomedical Research Centre, King's College London. This project was conducted with the aim of investigating clinical outcomes in mental disorders. Datasets reused in this project were produced at the South London and Maudsley Biomedical NHS Foundation Trust.

Data journey 5, *Understanding age inequalities and inequities in the cancer pathway* was conducted by a group of researchers working in the Bowel Cancer Intelligence (BCI) UK, a research collaboration hosted by the University of Leeds. This project was conducted with the aims of understanding the use of radical rectal cancer treatments and their related outcomes. Datasets reused in this project were produced at three different types of sites: NHS providers (acute hospital trusts, primary care trusts, and mental health trusts), NHS acute trust providers of radiotherapy in England, and NHS Hospitals.

Chapter 5. Socio-material drivers of patient data flows

5.1 Introduction

This chapter discusses the key factors that were identified in this research as drivers for the movement of NHS patient data through to universities to conduct research. The key factors identified here do not operate on their own, rather they interact and work together to drive the movement of patient data towards universities. This chapter introduces the factors identified and explains how the interactions between them created the conditions for helping drive the movement of patient data.

The findings presented here are informed by interviews with researchers and data practitioners that are part of research teams at different academic institutions and that work with data generated by the NHS. These interviews were conducted in 2018/2019 and analysed using thematic analysis. Whereas the main source for this section is the aforementioned interviews, data policy analysis is also used as a complementary source to inform this work.

5.2 Advancing health research

Currently, the UK healthcare sector faces a crisis as more people are suffering from serious illnesses such as cancer and heart conditions and the resources available to address this crisis are insufficient. The NHS provider sector is at the present time struggling to deal with increased demand as a result of a growing and ageing populations, and increasing costs. Funding for the NHS over the last decade has not been enough to cope with rising needs for healthcare. At present time the NHS does not have enough equipment or staff to meet the needs of the UK population (Charlesworth et al., 2019). According to a report from the Nuffield Trust, it is calculated that to be able to address the needs of an "ageing and growing population the NHS could need another 17,000 hospital beds by 2022" (McKeon et al., 2014, p. 2). Additional challenges for the NHS exist due to changing needs in healthcare, such as the rise in cases of antibiotics resistance, diabetes, and obesity (McKeon et al., 2014; NHS, 2019). The NHS itself and other governmental bodies have acknowledged the challenges aforementioned (NHS, 2014). Several participants of this research expressed wide awareness of this issue, and reported concern for this situation and commented that there is a need to find ways to address this crisis. The majority of participants expressed the belief that by doing research reusing patient data, they can contribute to addressing the current crisis at the NHS and ultimately benefit patients and the general population.

5.2.1 The desire to improve care and health outcomes

The objectives of the research projects in which interviewees of this study were involved at the time the interviews were conducted were very different from each other. For example: improving the diagnosis and treatment of stroke; trying to find ways to predict outcomes in psychotic disorder; and exploring long-term outcomes of urinary tract infections in childhood. Independently of their specific objectives, the majority of interviewees across all data journeys expressed that their goal was to contribute to advancing research by developing innovative treatments, and finding more effective and efficient ways of addressing health challenges:

> Better care, trying to fight [for] a population that is living longer and becoming unhealthy and we don't have the money to cure everybody sadly. Resources are finite, so you have to fight. So it's trying to find ways of making treatments more efficient or more things being done with the same amount of money. It's trying to deal with a population that is getting more and more overweight and that makes people more likely to get really nasty diseases and increase their health risk etc. So you got this combination of us trying to improve techniques and do things better for people fighting against the fact that the population is getting more and more unhealthy (S4-LS-2)

As with this participant, other interviewees similarly reported that their main goal when conducting research with patient data produced in the healthcare sector was to contribute to improving patient outcomes:

> better patient outcomes I think that's at the heart of this research, we want to improve things for patients, that is probably the main thing (S4-SS-4)
>
> In a university like here [the goal] is heavily around improving patient outcomes (S4-NS-3)
>
> Well there is a crisis in healthcare isn't it… we need to find out how to do things better... there are limited funds, even in the time without austerity there is a limit for budget... we have an aging population, we have more people comorbidities, that has never been faced before (S8-RN-1)

The desire to improve health and care outcomes has led researchers to look for opportunities to get access to patient data and conduct research to meet these objectives. They believed that given the fact that the NHS resources are limited, for them it is crucial to conduct research reusing patient data to help address the health issues of the UK population.

5.2.2 Framing their values against commercial interests

Some participants compared the drivers of academic research with those of industry, more specifically with those of pharmaceutical companies. Participants often evoked a form of virtue ethics, which suggest that to some extent, the actions of this community might be driven by moral concerns or by trying to pursue something good (MacIntyre, 2007; Powell, 2019). In interviews, participants often expressed that they were working with patient data because that would bring benefits to the patients and to the general population.

On the one hand they tended to adopt a posture stressing that their teams were committed to working to high standards and conducting rigorous studies to be able to deliver excellent research. For example, one participant explained that:

> I think everyone in the team genuinely share that feeling of wanting to do this to generate some sort of useful output and improve things for patients. And I think we are actually quite driven by the prospect of that and because of that I think everybody wants to do everything very rigorously and to have quite high standards in the analysis and the work we are doing. (S4-YS-5)

On the other hand, they argued that whereas at the heart of academic research projects lies a genuine intention to make contributions to improve the health and care of the UK population, pharmaceutical companies' main interest is commercial gain. For example, one participant commented:

> My understanding is that pharma companies are gonna be interested in commercial interest whenever they use datasets. On the other hand obviously we are not lucrative institutions...we are not a private organisation and the only thing that we are trying to do, our only concern is sort of guide and optimise antibiotic prescribing... and patient care without any commercial interest at all. (S1-LR-2)

University-based researchers seem to hold the belief that their values and motivations should be solid justifications for allowing health data from the healthcare sector to flow to them. On the other hand, they seem to believe that since researchers at pharmaceutical companies do not share their values and motivations, they should face more restrictions to access patient data.

According to some participants, the main driver of pharmaceutical companies is to make profit rather than improving health and they find this problematic:

> I suppose I am a bit sceptical about things like for profit companies, pharmaceutical companies having access to the data because that seems more like a case of profiting rather than trying to improve healthcare and patients' wellbeing and making sure there aren't any inequalities (S6-HM-3)

One of the main concerns expressed by participants in relation to pharmaceutical companies was that because their main interest is to make profit, they suspect that some of these stakeholders would be willing to misuse data and manipulate findings to take advantage and use the results of research for their own commercial advantage:

> Obviously that [using patient data] is lucrative to them, because they hope to develop a drug and they control the setting of the clinical trial and probably they can alter the results in their favour. Obviously they are gonna use that dataset for commercial interest. (S1-LR-2)

Another participant who expressed a similar idea to the one presented above showed concern about the potential misuses of data by pharmaceutical companies. However she also clarified that she is not completely against allowing pharmaceutical companies to access patient data, but stressed that if data is to be available for this type of stakeholder it is necessary to have adequate mechanisms in place that prevent the unethical use of patient data:

> there is huge business interest in patient data, particularly around drug companies, they are really keen to get into this data and I am a bit worried about that... so you can use the data in spurious ways to get the answers that you want, so you know "my drug is better than their drug, look these patients have better survival". Well is that because your drug was given to people who were younger and fitter and we can adjust for that? So, that is a big issue. I want everybody to have access to it, and I want access to the data and I want them to be used but you got to stop people playing nasty, you know manipulating it to get the end that they want. (S4-LS-2)

Having discussed the contrasting objectives of academic and pharmaceutical research, some interviewees argued that since the objective of academic research is to advance health, they should be granted access to patient data almost without any constraints as long as they are able to prove that their research is to be conducted in the public interest. This reflects the evocation of what Fiore-Gartland and Neff (2015) have defined as the transparency valence, in this context, the valence came to the surface when researchers talked about the benefits of sharing data generated in the healthcare sector with university-based research teams. When people evoke the transparency valence, they talk about the expectation of seamless flows of data. As one participant expressed:

> I suppose I am thinking of academic researchers where the aim is not commercial gain, the aim is to improve the health and wealth as a nation and NHS services, all those kind of things I feel like as long as you can demonstrate that is why you need the data and how you are using it then that should almost be completely granted with very little problems. It would seem sad not to allow academic researchers that perhaps are interested of the prevalence of a condition over time or whether the introduction of a new system has changed outcome over time... to not be able to really access that data. (S4-AS-1)

This participant further added that they feel uncomfortable with the possibility of pharmaceutical companies accessing patient data in the same way as academics, mainly because their driver is not to conduct research in the public interest:

> I am not comfortable with the idea that commercial companies or pharma companies could just equally access it in the same way [as university-based researchers]. I feel like the bar should be set higher when it is not for just public gain necessarily, which is for commercial gain which may have some public benefit, I think that is different (S4-AS-1)

This quote highlights that what seems most problematic for some participants is the idea of profit with patient data. As observed further in this chapter, the use of data for profit can also raise concerns among citizens. Moreover, whereas some participants were against facilitating access to patient data to pharmaceutical companies, others expressed a different view:

> In the academic context the motivation for doing research is to produce outcomes which benefit the patients. At the end the benefits go to the population, but in pharma companies it is kind of grey. But it is really difficult to know the exact reasons of pharma companies for using patient data... We need to recognise that pharma companies are also providing solutions to health problems, so even if they get profit, they are developing stuff that at the end of the day is useful and goes back to population, to the public. So, it is difficult to decide if this is a fair use of data or if it is not because you know as a researcher you can make a contribution but maybe on a smaller scale. (S7-GM-1)

This participant, who in previous projects worked in collaboration with private companies found it difficult to decide whether it was a good idea or not, to give access to patient data to pharmaceutical companies. He thought that whereas the goal of academic research is always to contribute to advance health, pharmaceutical companies' objectives sometimes become blurry because of their inherent commercial interests, which makes it difficult to evaluate whether their intended uses of patient data are fair or not. Despite the difficulty in fully grasping the intentions of pharmaceutical companies, some participants recognised these companies contribute to advancing health and to developing solutions to health problems and have the possibility to make contributions at a larger scale compared with university based research.

This framing of university-based researchers as ethical and well-intentioned people versus pharmaceutical-based researchers as profit driven seems to have the objective of creating a cultural construct that would allow for data to flow with fewer frictions to academic settings. This self-identity seems to be projected out with the expectation that it would enable them to have access to patient data with little or no constraints. Further exploration is needed to understand whether this has an impact on shaping data flows

.

5.3 The potential of routinely collected patient data

Findings of this study show that the ways in which university-based researchers value patient data and the expectations they hold about them play an important role shaping the movement of this type of data from the healthcare sector towards universities.

The belief that routinely collected patient data is a resource with great potential for advancing health research improving healthcare outcomes was identified as one of the key drivers for the movement of patient data generated within the healthcare sector. The findings of this study suggest this belief emerged and has been enhanced by four key expectations that the community of university-based researchers hold about data, here the notion of data valences is used to describe those different expectations. Three of the data valences identified belong to the ones defines by Fiore-Gartland and Neff (2015), namely actionability, truthiness, and self-evidence. In addition to these, another data valence, different to those was identified and defined drawing from the findings of this research: the vanguard valence (described in section 2.6).

5.3.1 Data save lives: the promise of data

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, p. 1474) was identified in the discourses of university-based researchers. As explained above, participants talked about their interest in advancing research and improving outcomes for patients, and when they did so, they also talked about how this goal can be achieved. When sharing their thoughts about the path to follow to reach this big goal of advancing and improving outcomes for patients, they talk about the potential that they see in the data and the reasons that have led them perceive such potential. In interviews, the expectation that 'data is actionable' often emerged; participants tended to highlight that 'patient data can help to save lives'. The idea that data saves lives was express by participants across all the different data journeys explored, from early career to senior researchers. For example, one participant commented:

> I think also there is a genuine belief that we can improve lives by using data, you know. (S3-EN-1)

Another 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-GM-1)

This quote seems to suggest that the idea that "data can save lives" is a belief that is not only present in some teams of researchers, but rather is a belief widely extended across the academic community.

While a strong form of the actionability valence was identified in the discourses of the majority of participants, some expressed a more nuanced view. Rather than saying that data on their own are able to lead to change, some participants merely noted that data can be a helpful resource with value and a great potential of help informing the development of policy and medical interventions, but not the only one needed to make a dramatic transformation in the healthcare sector:

> And I think also the potential that it could help develop interventions, clinical prediction tools which could be applied in the real world, I think the possibility of this. (S5-IN-1)

> Because I think there is value in the information that patient records contain. In my understanding and analysing all that information can be a really powerful guide to policy recommendations. (S1-LR-2)

The notion that "data save lives", which as stated above, evokes the actionability valence (Fiore-Gartland & Neff, 2015) was alsoidentified in public discourses, presentations or written pieces from senior staff representing key organisations in the UK healthcare sector landscape.

This is relevant in the context of this research because it reveals that the expectation that data is actionable, is prevalent not only within the community of university-based researchers, but also in the discourses of other key organisations within the healthcare data landscape. For example, in the policy conference 'The future of digital technology in the NHS - Big Data and AI, efficiency and outcomes, and addressing concerns' held on the 23rd of November of 2018, Daniel Ray, the Director of data of NHS Digital, opened up his intervention highlighting that:

> So our organisation, NHS Digital, indeed the whole health system role, is to harness the power of information and technology to make health and care better… we need to harness the power of data (Westminster Health Forum, 2018, p. 3)

More recently in the NHS Digital Transformation blog by Tom Foley, Senior Clinical Data Lead for Data at NHS Digital, on March 11th 2020 called "Data save lives", the power of data to improve health is highlighted:

> The NHS has some of the richest health data assets in the world. These are used by a range of people – researchers, doctors, government, bodies, charities – to improve our health. People don't always realise that getting the right data, in the right place, at the right time, can have as much impact as a prescription or a medical procedure. (Foley, 2020)

In a recent public appearance, Matt Hancock, the Secretary of Health also brought to the table the argument that "data save lives". During the opening keynote of the Founders Forum Health Tech Summit as part of London Tech Week, Hancock pronounced the following words:

> 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)

Nick Hirst, chief information officer at the National Institute for Health Research (NIHR) Clinical Research Network, in a comment published in the National Health Executive website on November 22th 2017 expressed that his organisation was working towards the goal of saving lives through data. In this comment he highlights that their end goal is to:

> Use the power of data to save lives and improve the health and wealth of the nation. (Hirst, 2017)

As will be discussed in more detail in Chapter 8, findings of this study also suggest that research using large datasets of routinely collected patient data is welcomed and endorsed by health research funders and policymakers in the UKConsidering the findings presented above it can be suggested that this shared discourse that "data save lives" which alludes to the expectation of the actionability of data might help to drive the circulation of patient data to academic research groups. Research groups might use the claim in funding applications and data requests that their main objective is conducting research to save lives. This argument is likely to be welcomed by UK health data research funding and policymaker bodies given the fact that they have openly expressed support for, and communicated that they value, data driven research.

Not only this, they have also promoted it and undertaken efforts to provide material resources to support data driven research.

5.3.2 Patient data as an accurate representation of the population's health

Two other expectations that were identified in the discourses of university-based researchers were the "truthiness" valence and the "discovery" valence. The truthiness valence, illustrates how people expect data to comprise a single, direct, objective representation of a measured reality, while the discovery valence depicts how people expect data to be the genesis of discovering phenomena, issues, relationships, or states that otherwise would remain unknown (Fiore-Gartland & Neff, 2015) It made sense to talk about these two valences in the same section, because in some instances they appeared together in the participants' accounts.

The majority of participants across all the data journeys explored expressed the belief that patient datasets generated in the UK healthcare sector are unique in the sense that they are better and richer than any other type of dataset. For example, one participant mentioned that:

> 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. (S6-AM-2)

It was observed that the truthiness valence (Fiore-Gartland & Neff, 2015) was present in the discourses of university-based researchers. This was evident in that a key reason for why participants believe that routinely collected patient data is a unique resource is because for them, unlike clinical trial data, routinely collected patient data presents an accurate representation of the 'real world':

> [Routinely collected patient data] just gives you a whole picture, real life, real world examples. It can give a different insight to clinical trials. (S6-HM-3)

This notion that data are accurate representations of the world has also been observed in other contexts and already contested by others in the CDS field (Gitelman & Jackson, 2013; Kitchin & Lauriault, 2014).

In addition to expressing the expectation that patient data can offer "a picture of the real world", participants of this research also highlighted that there are insights that would remain uncovered if they would not have access to this type of data. This revealed the presence of the discovery valence, which as explained before, describes the expectation that data can be the genesis of discovering phenomena, issues, or relationships that otherwise would remain unknown (Fiore-Gartland & Neff, 2015). For example, a participant expressed:

> I think there are lots of things you would not discover if you don't have access to these data. (S7-GN-2)

A number of participants across all the data journyes explored commented in relation to the potential of patient data that certain research questions could only be answered using routinely collected patient data. For example, a researcher who participated in the Building Rapid Interventions to Reduce Antibiotic Prescription project, (Data Journey 2), aimed at understanding the drivers of antibiotics prescription and their effects using routinely collected patient data explained that this study allowed them to understand important aspects of antibiotic prescription such as how antibiotics affected specific ethnic groups, and measure the rates of antibiotics prescription in areas with different rates of deprivation. She pointed out that without these routinely collected patient data they would have found it very difficult to find out answers to some of their research questions:

using these kind of dataset you get the real world view... this data is actually looking at who are being prescribed, what kind of dosage have been prescribed and it would be really hard to look at those questions without this dataset. (S6-AM-2)According to the majority of participants across all DJs, unlike routinely collected patient data, clinical trials data is not representative of what happens in real life. This is because often they trial medical procedures or drugs under 'ideal circumstances' and also because the patients recruited do not necessarily share the characteristics of patients that are seen by healthcare professionals in their day to day:

> we can do research with data which is representative of the real world in a way which for example randomised control trials often not, because the type of patients that they get recruited are not representative of the patients that we see in everyday clinical practices. (S5-IN-1)

Fiore-Gartland and Neff (Fiore-Gartland & Neff, 2015) observed that the discovery valence follows the logic that finding patterns in data is equal to knowing or understanding patterns in life on cellular, individual, or population scales. In this research, something similar to this was observed, in this case, university-based researchers across all data journeys tended to perceived that finding patterns in data was the same as understanding patterns in the health of the population.

The two expectations discussed here, that 1) patient datasets provide real world examples, and, that 2) finding patterns in data is equivalent to understanding patterns in the health of the population might help drive data towards university-based research groups. For researchers and data practitioners across all data journeys, having access to what they perceive to be the whole picture of a phenomenon seems more appealing than just studying a sample. The expectation that with patient datasets it would be easier to answer their research questions, and that these datasets can act like a window to observe what happens in healthcare settings, seems

to be a key motivator for university-based research groups to seek opportunities to make data flow into their labs so they can conduct research with this type of data.

5.3.3 Routinely collected patient data as a resource that provides better insights

A form of the self-evidence valence was identified in the findings of this work, this valence depicts the expectation that data are premade, therefore they do not require work or interpretation (Fiore-Gartland & Neff, 2015). Participants across all data journeys evoked the self-evidence valence, however identification with this valence was not as strong as described by Fiore-Gartland and Neff. They argue define the self-evidence valence as the firm neglect of the ideas thatintervention of people is needed to control, organise, and structure to transform data into something meaningful. and completely overlooking the fact that selecting what data to collect, making sense of the data, and arriving at a conclusion needs a theoretical foundation, a frame, or a hypothesis. In this study, participants had a better recognition of the complexities of working with data. They talked about their work with data, and when doing so they explained the several steps in which they engage before drawing conclusions from the data collected. They talked, for example, about the need of cleaning and organising data, running different rounds of analysis before being able to yield results, and the challenges that they experienced working with data. These challenges will not be detailed here as they are a matter of discussion in chapter six, where the frictions for the movement of data are discussed. Interestingly, while participants talked about these different stages of their research process, and the difficulties that they experience in each of them, when explaining why they perceive patient data as a resource with great potential, the contradictory expectation that data are self-evident emerged.

=

A number of participants across all data journeys expressed that they see a great potential in routinely collected patient data because of its large scale. According to the majority of interviewees, decisions should be made based on data; thus having access to larger amounts of data allows them to conduct a better analysis and offer more accurate answers to research questions. As one participant reported:

> You remove guess work, you know in science you should make every single decision based on data. The less data analysed, the more you are guessing in any decision made and that's kind… the more information you get the more informed every decision is ever made. (S3-EN-1)

Another participant expressed a similar view to the one from the quotation above, highlighting in addition that the level of insights that can be gained with this type of data cannot be obtained from analysing other type of datasets:

> There is value... having data on a large scale about people can answer or can do lots of things, it can be used to answer questions that are relevant to society in a way that you can't with other datasets. (S6-NM-1)

When talking about the potential of doing research using routinely collected patient data, participants who in the past had the opportunity to work with clinical trial data often compared the insights they can gain using routinely collected patient data with the ones that can be obtained from data gathered via clinical trials.

An example of this is depicted in the following quotation, in which an interviewee comments that routinely collected patient data allows a broader scope than clinical trials data (which is more limited):

> So it seems you can do much broader things by using secondary used data than if you were trying get a study where you have fewer people and do lots of different things you would not be able to do things on such a scale if you were not using these datasets. (S6-HM-3)

To exemplify the perceived advantages of large sample size of routinely collected patient data, a member of the *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders* (Data Journey 4), compared a research project in which they were currently involved, using routinely collected patient data to examine psychotic disorders with a study with similar goals, conducted in the past not using routinely collected patient data. This interviewee highlighted how using routinely collected patient data in their research allowed them to get richer and more detailed information for their study than the other research they recalled:

> It's a large sample size and we have quite rich clinical data about the presenting symptoms that people present with and in a way that we just haven't had available before at a population level. And I think previous research developed clinical prediction tools for relapse have not had access to this fine grain detailed, not detailed information about the clinical presentation or the nature of illness that people experience. (S5-IN-1)

In this case, the self-evidence valence is perceivable in that the participant assumes that a dataset offers comprehensive details about an illness.

5.3.4 As explained before the majority of researchers and data practitioners across the different data journeys explored hold the belief that patient data are accurate representations of the reality, and that is already a key motivator to conduct research with this type of data. The belief that large datasets provide better insights seems to be an additional incentive for researchers and data practitioners to conduct research with patient data and to seek opportunities to continue doing it. It would seem that from their perspective, the more data you have, the more you know. This has led university-based researchers and data practitioners to see large patient datasets as better data than other types, such as clinical trial data. **Non-intrusive analysis: 'Do no harm'**

Some participants explained that another advantage associated with using large datasets of routinely collected patient data, and a reason for why they consider such datasets as resources with great potential, is that these datasets enable researchers to study certain aspects of an illness or particular conditions without causing any harm. This view was expressed by participants that before engaging in data reuse projects, had the opportunity to work with clinical trial data (e.g. researchers working in Data Journey 3 which explored urinary tract infections in childhood). According to these participants, reusing routinely collected data is particularly useful when studying certain conditions or illnesses that cannot be studied in another way without endangering or posing risks to the individuals involved.

For example, a researcher who at the time the interviews were conducted was participating in a project aimed at investigating antidepressant-related issues explained how reusing patient data has opened up for them the possibility of understanding certain features of antidepressant use in the population without asking individuals to participate in a clinical trial:

> With large datasets you certainly have the power to find all of these different associations without causing harm. We can identify particular issues or sort of create a decision aid which mean people are being prescribed antidepressant A rather than antidepressants B because they are already on drugs X, Y and Z. (S6-HM-3)

Similarly, another participant reported that routinely collected patient data allows researchers to study certain medical conditions suffered by groups of people that often are prevented from taking part in clinical trials due to the risks associated with this type of studies:

> and it was about the value and this kind of research compared to say, so the clinical trials data. They have massive gaps in terms of who they cover, who is included. Children are often excluded, elderly people, people who are the most ill, they are excluded from trials. (S6-AM-2)

For researchers who are familiar with the complexities of studying certain health conditions relying only on clinical trial data and the risks that taking part on studies of this nature represents for some individuals, conducting research with patient data is seen a viable alternative to explore critical health issues without harming the population studied. This factor has helped drive the circulation of patient data to university-based research groups in the sense that this seems to be an important motivator to seek opportunities to conduct research with patient data, particularly for those groups that have experienced challenges to explore certain health issues in clinical trials due to the potential risks for the groups intended to be studied (e.g. Data Journey 3 which explored urinary tract infections in young children and Data Journey 4, which investigated clinical outcomes in mental disorders).

5.3.5 Routinely collected patient data as an efficient and cost effective resource

According to participants across all data journeys, another reason why routinely collected patient data is considered a resource with great potential is because compared to using data collected in clinical trials or in other types of medical research studies, reusing routinely collected patient data results in a more efficient and less expensive way of gathering data for conducting research. For example, one participant who worked in *Understanding age inequalities in the cancer pathway* (Data Journey 5), a project that linked several datasets from four different types of NHS service providers explained:

> The potential of it. It's much more efficient, much more cost effective. it's crazy not to use it... we can use data to improve and make things move. (S4-AS-1)

Participants explained that reusing patient data is less resource intensive for them. Since this data is collected at the point of providing care, they do not need to invest time and financial resources in gathering these data as they would have to if they were doing a clinical trial:

> If the data is there and obviously it prevents you from going and then collect more information, spend resources like time... yeah it is just a very efficient way of yeah, using data. (S2-FA-1)

According to interviewees who before start working with routinely collected patient data, reusing patient data is not necessarily cheap as they have for example to pay fees to the NHS or data providers for accessing data or for data linkage services; however, they recognise that reusing data is less expensive than producing their own datasets. As one participant who was part of the *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders* team (Data Journey 4) commented:

> Well I mean, I think the fact that is already there means that the cost is relatively low. I say relatively because it still expensive but nowhere near as expensive as having to develop your own curated dataset

> where you interview individual people and you get them to fill questionnaires and things like that is incredibly resource intensive but this data already there routinely. (S5-IN-1)

Similarly, another participant reported:

> So if you had a cohort study that you have to build yourself and take all that time to build up and it would be super expensive. (S6-AM-2)

University-based researchers and data practitioners across all data journeys believe that a great advantage of patient data research is that they do not need to conduct the data collection themselves. This belief has led them to prefer seeking opportunities to perform this type of research over qualitative studies or clinical trials, based on the argument that this offers them the possibility of spending less financial resources and time. In addition to this, the efficient use of resources is highly valued across universities; therefore conducting research with fewer resources is often welcomed and encouraged.

5.3.6 The demand to exploit data

Participants across all data journeys explored pointed out that at the present time it is possible to access more patient data than ever produced in the healthcare sector. As previously discussed, university-based researchers and data practitioners believe that this type of data offers a number of advantages, which are that they are an accurate representation of the world, that because of their large volume they provide better insights, that they offer the possibility of exploring health issues without causing any harm, and that they are cost effective and efficient. Given that they believe in these advantages, for them the most sensible thing to do is to exploit it for research purposes and overcome the underexploitation of this abundant resource.

In summary, according to participants, the great potential of patient data results from the combination of four key factors: it can offer large samples of the population intended to be studied, requires fewer resources to be collected, has less restrictions in terms of data included and what can be studied, and offers real world examples.

As one participant commented:

> The fact that we have this data now, we have electronic health records, so 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-IN-1)

Some participants commented that despite the NHS having produced and stored routinely collected patient data for many years, these large amounts have been barely used. This view is captured in the following quotation:

> Because I think there is value in the information that Electronic Health Records contain, for example Clinical Practice Research Datalink has been like collecting information at least since the year 2000, maybe 1997, that's a long time already, and that information has barely been used and analysed. (S1-LR-2)

Some participants also pointed out that for them it seems counterintuitive to not use these large amounts of data that have been already collected and are stored "waiting to be reused":

> Data is sitting there so it makes sense to use them. (S6-HM-3)

> I mean the data already exist so we should be using it to improve care, improve services, is like a no brainer. (S2-FA-1)

According to some researchers in senior positons with several years of experience working with patient data, despite the NHS generating large amounts of data that could be used to conduct research, better understand illnesses and inform decision making to design adequate health interventions to help fight illnesses that are affecting the population, the NHS has been struggling to do this mainly because of a lack of resources. They also pointed out that it is because the NHS does not have enough resources to analyse the data they produce, that it is necessary to enable external stakeholders to reuse this data for research purposes. For example a researcher working on *Building Rapid Interventions to Reduce Antibiotic Prescription* (Data Journey 2) expressed:

> The reason it needs to be shared out and it needs to go for secondary use is because it doesn't get done in house. This stuff is just routinely collected and it just, I don't know the records just build up. They don't have researchers in house, that would be amazing if they did, but they don't and I don't think that's gonna come out of their budgets. So your GP practice does not have a research side in it, they might have a data manager, but they are not the ones raising hypotheses and testing them. So it's got to be those questions come from outside and the expertise in order to answer them comes from outside. (S1-YR-1)

The notion that patient data is a highly valuable resource that should be exploited contributes to driving the circulation of patient data towards university-based research groups. Since patient data is perceive as highly valuable, university-based researchers are encouraged by their institutions, funders, data providers, and policymakers to conduct research using them. It is for this same reason that other types of research considered less efficient and expensive, such as a large scale qualitative research or a clinical trial might not receive the same support as health data research.

5.4 Technological advances

Technological advances were identified as key a socio-material factor that have helped to drive the reuse of patient data for research purposes outside the healthcare sector. According to participants of this study, the most important advances are the digitising of the UK healthcare system, a broad range of innovative analysis tools and techniques, and the development of

robust systems and mechanisms to protect confidentiality of patient data. How these three infrastructural factors contribute to driving health data circulation to university-based research groups is addressed in this section.

5.4.1 Digitising of the UK healthcare system

The general practice sector in the UK started its digitisation in the 1980s and by the mid-2000s it was almost completely digital. Thanks to this, primary care patient data has been available for reuse for a long time. For example, Clinical Practice Research Datalink has been collecting de-identified patient data and supporting research studies for a long time now. CPRD is a not-for-profit research service of the UK government (Department of Health and Social Care, 2011), previously known as General Practice Research Database (GPRD), which has been operating as a data custodian offering de-identified primary care patient data for over 30 years (Clinical Practice Research Database [CPRD], 2018).

CPRD in its current form, 'Clinical Practice Research Datalink', was introduced in April 2012 with a key feature being the introduction of linked healthcare records (Department of Health and Social Care, 2011). CPRD is sponsored by the "Medicines and Healthcare Products Regulatory Agency and the National Institute for Health Research (NIHR)" (Clinical Practice Research Database [CPRD], 2018, para. 1).

However, the digitisation of the healthcare system has not advanced at the same pace in other areas of care and services. That is the case in secondary care, where digitisation is still a work in progress as the National Programme for Information Technology, created in 2002 and aimed at digitising secondary care, was cancelled in 2011 having failed in achieving almost all of its goals. Things have been changing in recent years as the government, recognising the need for a digital transformation of the healthcare sector, has undertaken efforts to make this possible. One of these actions was, for example, that in 2015 the Treasury provided funding of more than £4 billion to accelerate the digitisation of the NHS.

Thanks to the advances in digitisation, more patient data in electronic format has become available for conducting research outside the healthcare sector. This, according to some interviewees, has facilitated the work of academic researchers in reusing patient data in research:

> Increasing digitisation of healthcare information, that's been quite helpful. (S3-EN-1)

This does not come as a surprise, in fact other studies in the past have highlighted that one of the factors that contributes to fostering data flows is the availability of records in digital format (De Roo et al., 2016).

Findings of this study suggest that the availability of data helps to create a desire to request to use it and therefore drives the circulation of the data. According to some participants with several years of experience working with patient data, the fact that the shifting materiality of patient data from analogue to electronic format had become available motivated researchers to request access to make use of it. As one interviewee who worked on *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders* (Data Journey 4) tells, in recent years more and more members of the research community began to show interest in accessing patient data for conducting research:

> There has been a lot more interest in this [using patient data for research in academia]. When I first started working on these datasets not many people have heard of the concept of electronic health research and now everyone in this institute would have heard of it and quite a lot of people if not sort of directly or indirectly been involved in some research using electronic data and I think more people want to do research projects, so the demand has increased. (S5-IN-1)

The above quotation highlights the fact that in recent years the reuse of patient data started to gain traction. Whereas in the past university-based researchers were unfamiliar with the practice of conducting research reusing electronic patient data, now many academics are familiar with this practice as they themselves have participated in research projects using this type of data, or if not, at least they know someone who has done it in the past or is doing it.

Technological advances have helped generate a growing demand for patient data for secondary purposes, which has helped drive data towards university-based research groups. The findings suggest that a repeating cycle has emerged where the more technology developments allow the data to be available, the more demand increases, and the more technology is developed to enable more data to be available.

5.4.2 The opportunity to apply innovative techniques to analyse data

The accounts of some interviewees seems to suggest one factor that has fostered the growing interest among university-based researchers in reusing patient data for conducting research is the desire to work with data using a broad range of tools and techniques that are available at the present time. Thanks to the advance of technology they have been able to innovate and explore what new things they can do using new data analysis techniques. The majority of participants across the different data journeys explored expressed that they felt attracted to

work in projects that involved the reuse of patient data because of the possibility of exploring the potential of new technologies. For example a member of the *Using technology and data to improve the diagnosis and treatment of stroke* team commented :

> I wanted to do something more novel with the professors who are trying to design their own research and design their own data model 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-YR-1)

Some participants, mainly those working on large-scale projects seemed to be particularly interested in linking datasets from many different sources and in extracting multiple "chunks" of data from datasets at their convenience For example a researcher working on Data Journey 5, a project in which several datasets extracted from six different types of NHS service providers expressed:

> The madness of it...I can link all these data together and I can get really good intelligence. (S4-AS-1)

Similarly, another research working on this same data journey commented:

> I find it really interesting that with data you can cut and slice data however you wanted. (S4-NS-3)

The majority of participants across all data journeys highlighted on multiple occasions during the interviews that technological advances have increased their capacity to exploit data. When explaining how these advances have changed the way in which they do things, they provided explanations as the following:

> …and the power of computer and machine learning, that just be able to increase our capacity to analyse these datasets and to see more and get more out of the data. (S3-EN-1)

The above quotations suggest that the potential to extract more informational value from data and enhancing the analysis capacity are perceived as important benefits brought by technological advances.

Participants talked with enthusiasm and excitement about the diverse ways in which they can analyse data, giving detailed accounts about how a broad range of analytic techniques worked. Some interviewees talked in detail about their work doing natural language processing, cluster analysis, pattern mining, neural networks analysis, and developing algorithms. When talking about these techniques they also explained how they navigate these complex analysis processes using a number of different software and packages such as R or Python. For example, one participant, when asked to provide an example of an innovative technology they have used or are using at the present time, gave the following account:

> I am now using Natural Language processing techniques… I am applying is text mining software to the free text on the notes to put out clinical information such as what symptoms someone has been presented... and develop prediction model to predict relapse. (S5-IN-1)

Whereas interviewees talked about a broad range of different techniques they have used to analyse data, the technique that seemed to generate more enthusiasm and excitement around participants was the predictive analytics. Yet, the excitement expressed for predictive analytics is not necessarily prompted by the results obtained using this approach. For example, a participant working in Data Journey 4, aimed at predicting outcomes in psychotic disorders, showed excitement for predictive analytics despite not knowing yet if successful results would be achieved. They commented:

> ...to try and develop a prediction model to predict relapse and the other part which essentially will only is to see if it is possible to predict relapses... now we are just trying to develop clinical prediction models, refining the data extraction to see if it's possible to accurately predict relapse so that stage I am at the moment. 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-IN-1)

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. So I'm doing cluster analysis, I'm doing pattern mining and models so I take people who have dementia and look back 20 years and I look at any signs, symptoms, medication, procedure, literally anything and I try and find patterns and sequences over the whole 20 year period before they are diagnosed to see if I can define. (S3-EN-1)

Both of the accounts presented above are from researchers who, at the time the interviews for this research project were conducted, were undertaking the described projects. Whereas both of them were at an early stage and did not know if they would be able to develop the prediction models intended, both of them thought it was worth exploring this possibility and felt excited about the prospect of achieving it.

Using the notion of data valences (Fiore-Gartland & Neff, 2015), I propose that when researchers talk about the opportunities of applying innovative techniques to analyse data, they evoke what I define as the vanguard valence. The notion of the vanguard valence aims to illustrate how people perceive conducting research with patient data as the most innovative, cutting edge,way of exploring health issues. Participants often highlighted that this type of research positions them as part of a group that is breaking with old ways of doing science.

The excitement for applying 'innovative techniques' to conduct research using data helps to drive the circulation of data to university-based researchers. In the first instance, it would seem that this excitement motivates them to come up with ideas for research proposals through which they have the potential to be recognised as pioneers in applying certain analysis techniques, despite not having the certainty that such innovative methods would help them to address their research questions. Moreover, proposing research using 'innovative' techniques helps them to obtain

support from funding bodies, which have expressed that they value innovation and data-driven research, and have offered to support and provide funding for conducting this type of research.

5.4.3 Development of robust secure storage systems

Technological advances have also facilitated the reuse of patient data in the academic context because they have provided the means for handling, storing and transferring patient data in a more safe and secure way.

Data safe havens

As commented by participants, technological advances have made possible the development of data safe havens to store patient data, which has facilitated researchers in conducting research projects while ensuring data is secure. Some participants talked, for example, about how the development of an innovative secure data analytics facility in their institution had enabled them to maximise data security, maintaining data confidentiality, integrity and availability. This is the case of researchers working on Data Journey 2, who stored all the data used for their research project in a secured safe environment developed by their institution. Talking about the benefits of safe havens, a participant commented that working within an organisation that had recently developed its own data safe haven using the most up to date technology allowed their team to access patient data in an efficient way and gave them the confidence that:

> The security around the data is appropriate, you know the data I have access is completely pseudonymised and secure. (S3-EN-1)

Participants working on other projects that lacked this facility had to make use of external safe havens to be abe to handle data in a safe way. For example, one of the researchers working on the project *Long-term outcomes of urinary tract infection in childhood (LUCI)* (DJ3) commented:

> We actually use a data safe heaven, so we work with the SAIL Databank in Swansea and they hold the data for us because they have secure environments…we make sure that the data is secure. (S2-FA-1)

It can be suggested that the existence and availability of data safe havens has helped to drive data towards researchers not only because researchers feel confident using them, but also because having this infrastructure in place is a factor that gives the NHS, policymakers, patients and the public the confidence that data can be shared with academic institutions safely and securely. These data safe havens are therefore an important part of the material infrastructure that enables data flows.

Automated anonymisation

When discussing the technological advances that have facilitated the reuse of patient data outside the healthcare sector in a secure way, apart from data safe havens, the other development that was highlighted by participants was the automated anonymisation:

> Creating and running algorithms that enable these datasets to be linked to create anonymised versions that can be made available for research. (S4-AS-1)

Participants see this technological capability as a facilitator because thanks to this, anonymisation can be done without human intervention, as a participant explained:

> Because now de-identification process is automated, database goes through a pipeline everyday where new data come through and automatically deidentify into the database, and that deidentification is a machine learning based system. (S5-IN-1)

Whereas many participants talked with enthusiasm about the advantages of the automated anonymisation, some participants were more cautious and sceptical when talking about these developments. Researchers who have detected flaws in the automated anonymization processes in projects in which they have participated highlighted the fact that this, as other technologies, are not perfect solutions and are susceptible to failure. For example, a researcher working on *Using technology and data to improve the diagnosis and treatment of stroke* (Data Journey 1) pointed out that organisations cannot rely blindly on automated anonymisation processes; rather, they need to apply rigorous monitoring procedures to ensure the anonymisation has been carried out successfully:

> The organisation holding the source data doing the checking before it sends it off to some other organisation... that doesn't seem to happen, they just send it, and there's like there's no checking. I really wish there was checking because… is a huge risk for that organisation if something goes wrong, if they sent the wrong file, if they send it to the wrong place, they include the wrong column, potentially that's a huge breach that could cost them a lot of money. (S1-YR-1)

The possibility of efficiently applying anonymisation processes to datasets helps driving data towards researchers. Despite some university-based researchers seeming to be sceptical about the use of automated anonymisation processes, the use of these processes is generally seen by as a good method to ensure individuals' privacy in datasets is protected. Therefore, applying anonymisation techniques gives researchers the confidence that they have appropriate methods in place to ensure sensitive data are protected. In addition to this, data providers' utilisation of an anonymisation process is also perceived as an appropriate approach to ensure privacy.

5.5 Advancing academic careers

The interest of researchers in using patient data to conduct research has led to the creation of socio-material drivers for the flow of this type of data towards university-based research groups. These groups expressed that they felt motivated to reuse patient data because this provides work opportunities for them and allow them to meet the demands of academic institutions regarding the publication of journal papers. Working with patient data and meeting institutional demands of creating new material resources (papers) leads them to obtain social (prestige) and material (financial) rewards.

5.5.1 Creation of jobs for researchers

When asked about the reasons why they are interested in reusing patient data, one of the reasons mentioned by participants was that doing research with this type of data is creating jobs for researchers:

> [Patient data] is just providing a job for researchers. (S4-LS-2)

Similarly, other participants explained that they were aware that reusing patient data could potentially benefit them. For example, another participant mentioned:

> I do recognise that as a researcher doing this work does ultimately benefit me more than not doing it if I was in the same job. (S4-YS-5)

University-based researchers recognise that the availability of data helps in creating and sustaining job opportunities for them. For them, if data stops flowing, jobs are at risk of disappearing. This might be a factor motivating them to continue making efforts to obtain access to new pockets of data to be able to extend existing research projects or conduct new ones.

5.5.2 Meeting publishing demands from academic institutions

Another key benefit that university-based researchers across all the different data journeys explored perceived was that they could report the results or methodologies of their research in academic papers, which could help them advance their academic careers:

> From a cynical perspective, academics are interested in advancing their careers, so they want those high profile papers, they want to stand in front of their peers at international conferences, be lauded for their intellectual bigger insights. (S4-LS-2)

The view that papers and publications are very important was shared among participants of this study, from early career researchers to senior academics. For example, an early career researcher commented:

> I'm quite junior academic type staff so kind of publications and papers is a big thing career wise. (S4-YS-5)

Similarly, a senior academic, who is a lead researcher on a project, commented:

> Obviously the researchers benefit because you know if you are an academic you need publications. (S7-GM-1)

Doing research that can lead to the publication of academic papers seemed to be quite important for participants of this research as they feel that producing papers can help them to maintain and protect their position in academic institutions:

> Any higher education institution always gonna require papers out, you know, you're protecting your position. (S4-LS-2)

Academic institutions expect researchers to produce papers in order to fulfil their institutional objectives such as earning prestige, reputation, and attracting more students. One of the factors used to evaluate the performance of researchers is their publishing record and their career progress is highly influenced by this; the more they publish, the more they can progress. Therefore, a factor that might help increase the demand for patient data is that conducting research with patient data provides researchers with material to produce papers.

5.6 Chapter conclusion

This chapter presented a number of social and material factors that together have driven the movement of patient data towards university-based research projects or groups. The factors identified here do not work independently as on their own they cannot shape data flows, rather the interactions between them, strengthen one another, and when this happens they can influence the ways in which data flows.

University-based researchers have expressed that they wish to contribute to improving the health and care of the UK population. This wish has led them to seek opportunities to engage in health research because they believe that through this practice this wish can become a reality. The desires of a community certainly play a role in shaping the flows of data, as Bates and colleagues (Bates et al., 2019) observed in their empirical exploration of meteorological and climate data flows, the desires and enthusiasm of key actors interact with the socio-material context to foster the movement of data. Similarly, this study revealed that university-based researchers across the diverse data journeys examined have embraced the belief that patient data is a resource with great potential, mainly as a result of the four key expectations they hold about them. To unveil these key expectations, the notion of data valences (Fiore-Gartland &

Neff, 2015), was used to explore the different expectations and assumptions researcher held about patient data. It was observed that participants tended to perceive data as 'actionable', that is to say that they expect that patient data can drive a change within a social setting or can be leveraged for action, to be more specific, one way in which this expectation was evident was in that participants often said that 'data saves lives'. As explained in this chapter, this expectation that patient data are actionable is also present in the discourses of key actors in the UK healthcare sector landscape, such as funders, policy makers and data providers. The 'truthiness' valence was identified in that participants tended to perceive data as accurate representations of the population's health and as resources that can offer better insights than other types of data, such as clinical trial data. A form of the 'self-evident' valence, which depicts that data are premade resources that require little work or interpretation was also identified. As will be discussed in more detail in Chapter 8, some of these views are shared by some data providers, policymakers and funders, which often have encouraged and support university-based researchers in conducting research with patient data.

Specifically, while it was observed that participants talked about the difficulties for cleaning, analysing and making sense of data, these difficulties tended to be obscured when they talked about the benefits of working with patient data as they tended to highlight that with patient data great results can be achieved with little work. The difficulties experienced by researchers when cleaning and trying to make sense of data and how this interact with the desire of researchers for reusing data because these are perceived as resoruces with 'great potential' will be explored with more detail in Chapter 6.

As observed in section 5.2.2, findings of this research show that across all data journyes explored university-based researchers seem to perceive themselves as a virtuous group of people driven by moral concerns (Powell, 2019) and have tended to frame their values against commercial companies. They have felt that the fact that they have this main objective should be a solid justification for allowing patient data from the healthcare sector to flow to them. Here we can see manifested another expectation of the members of this community, evoking the 'transparency' valence, which describes how people expect seamless flows of data across different contexts (Fiore-Gartland & Neff, 2015). In this particular case the university-based researchers across all data journeys explored seem to expect a frictionless flow of patient data towards them because they are an ethical and well-intentioned community. This framing of researchers as a virtuous group seems to be projected out with the expectation that it would enable this group to have access to patient data with little or no constraints. However, as

explained earlier further exploration is needed to understand whether this has an impact on shaping data flows.

Technological advances have helped to create a growing demand for patient data to be reused for secondary purposes, which has helped drive data towards university-based research groups. In the first instance, thanks to the advances in digitisation, more data in electronic format has become available for use outside the healthcare sector. This has motivated researchers to request to use these data in their research. It is well known that advances in digitisation is a factor that contribute to smooth the movement of data, this is something that other studies have highlighted in the past, including recent research in other context (De Roo et al., 2016) Findings of this study suggest that a repeating cycle has emerged in which the more technology developments allow the data to be available, the more demand increases, and the more technology is developed to enable more data to be available. This technologically rooted cycle is also reinforced and deepened through its interactions with the desires and beliefs of researchers as described above.The advance of technology has also made it possible to use new data analysis techniques. This has generated enthusiasm and excitement among university-based researchers, giving birth to another key expectation. It would seem that this excitement has motivated some researchers to develop research proposals applying novel data analysis techniques that could lead them to be recognised as innovators. To refer to this expectations, building on Fiore-Gartland and Neff's concept of data valences I defined this as the 'vanguard valence': a term that illustrates how people perceive conducting research with patient data as the most innovative way of exploring health issues of the whole population.

Technological advances have also enabled the reuse of patient data by researchers at universities because they have provided the means for handling, storing and transferring data in a more safe and secure way. Having this infrastructure in place and the existence of anonymization processes gives university-based researchers, data providers, funders, policymakers, patients and the public the confidence that data is being shared in a safe and secure way, and that privacy is being protected.

An additional factor that has helped to drive the movement of patient data towards universities is that university-based researchers have felt motivated to reuse patient data because this provides work opportunities for their groups and allows them to meet the demands of academic institutions regarding the publication of journal papers.

Chapter 6. Socio-material sources of friction of patient data flows

6.1 Introduction

The previous chapter identified the key socio-material factors driving the movement of patient data from patient records to researchers. This chapter presents the views of researchers and data practitioners at sites of data reuse in relation to the different challenges and issues that can emerge as patient data produced in the healthcare sector move from the site where they were originally generated (site of data production) to sites of data practice where they are reused for conducting academic research (sites of data reuse). This chapter is organised in three main sections: section 5.2 discusses the data sharing infrastructures and management; section 5.3, regulatory frameworks; and section 5.4, sociocultural factors.

6.2 Data sharing infrastructures and management

Data sharing infrastructures and management, including "technical infrastructures, data management practices, organisations and materialities" (Bates, 2017, p. 416), are a combination of interrelated social and material elements that can constitute a source of friction for patient data generated within the healthcare sector that works against drivers to get data moving (Bates, 2017). As discussed in the literature review, data sharing infrastructures are the physical and digital resources used to store, share and disseminate data across networks that are at the same time embedded within broader institutional scenarios (Kitchin, 2014). Researchers have observed that the complex constitution of data infrastructures reveals that they lack neutrality in the same way as data; they are dynamic, material manifestations of some broader movements of social and political nature (Furlong, 2011; Harvey & Knox, 2012). The key friction-generating factors related to data sharing infrastructures and management identified by participants are explored in this section.

6.2.1 Fragmentation of data creation

Whereas significant amounts of patient data are generated, these data are held by different organisations within the healthcare system that have different ways of managing data (Hall & Pesenti, 2017); this can make data flows in the UK healthcare sector difficult to trace and understand. The messiness of patient data streams was perceived as one factor that hampers the movement of patient data from the healthcare sector to academic institutions to conduct research. Participants across all data journeys argued that not having all patient data available

in a single source can slow down or block the movement of data generated within the healthcare sector. They believe that it would be much more convenient to have all data from a single patient recorded in a single healthcare record. This record, they argued, ideally should be used to capture patient data from primary, secondary and tertiary care produced at any NHS provider across the UK. As one participant said:

> We are not doing enough for sharing. Data is everywhere and the simplest case should be a single shared digital care record for each patient. (S5-IN-1)

In order to be able to access patient data, researchers first need to map the data flows to know how to access them. According to interviewees, data journeys across the healthcare sector are extremely complicated and difficult to follow, which creates a friction for data to flow from the healthcare sector to universities. This happens in part due to the national differences that exist in the way patient data are managed across the four nations that make up the UK. This is also because of the fragmentation of data within the same nation or even within the same site of data practice.

National differences

According to participants who have worked in projects that require data from different nations, one of the reasons why it is difficult to access patient data available across the four nations that make up the UK is that there is not a straightforward and easy way to map these data because each nation manages their flows of data independently (establishing their own guidelines of collecting, storing and sharing this data). For example, one participant that collaborated in the project Long-term outcomes of urinary tract infection in childhood (Data Journey 3), which gathered data of children from different UK nations commented:

> There is the Welsh organisations, then NHS Digital, the Scottish system… I think the information sits with each of them…so even just being able to make the data come from three countries is challenging. (S7-GN-2)

The national differences seem to be problematic because the collection methods can be different, the computer systems do not have a way to interoperate because it would seem that there is no organisational incentive to overcome technical barriers, and sites of data production based in different nations are not able share the data they produce with each other. This means that researchers intending to use patient data from UK nations need to deal with non-standardised data. In addition to this, they need to make sense of how data flows work and are managed in each country and follow independent procedures to access the required data. Researchers who used data from only one nation did not experience frictions generated by the

national differences described here, however this issue would limit the volume of data available for projects focused on exploring a rare condition or seeking to examine health outcomes among different countries or regions.

Fragmented data journeys within the same nation

In addition to the difficulty in mapping and accessing available data for conducting research due to regional differences, some participants also found it problematic that within one single region, patient data is disjointed. As explained in a previous chapter of this work, patient data can be recorded in varied ways. For example, primary care data is recorded in electronic patient records because the computerisation of general practices started in the early 1980s, whereas in secondary care it is still common to have handwritten records.

One participant who worked in the project *Using technology and data to improve the diagnosis and treatment of stroke* (Data Journey 1) explained the frictions they have experienced due to the data fragmentation within the same region. For this project , they were required to work with patient data produced by an ambulance service and patient data produced at a hospital. Since these data were stored independently and in different formats, in order to be able to undertake the project they had to conduct a series of technical arrangements to work with data from these different sources:

> So one dataset we've got recently…there was a problem of a missing column. First you need to link the data, the ambulance and hospital data and then you need to get a statistician to look at it to work out. (S1-YR-1)

The fragmentation of patient data not only complicates the process of understanding data flows in the UK healthcare sector. It also sets barriers for negotiating access to patient data, and brings technical complications if data from different sources, or recorded in different formats, need to be linked. This issue was more specific to those data journeys where data travelled directly from health providers, for example Data Journey 1, *Using technology and data to improve the diagnosis and treatment of stroke* (in which researchers obtained data from the North West Ambulance Service and Salford Royal Hospital), compared to data journeys where data came from intermediary sites such as NHS Digital, CPRD or SAIL Databank who work to overcome some of that fragmentation e.g. data journeys 2, 3, and 5.

Fragmented data journeys within the same site of data production

A number of participants reported having difficulties making sense of data flows of data generated within the same site of data practice, as even within the same site of data production, patient data are not always joined together. They commented that it is common to see that data

cannot travel across the different departments of a single site of data production because not all departments use the same computer systems or data collection tools and strategies. As one participant explained:

> "They have lots of different clinical systems that are all running but they are not brought together…the hospital does not bring it together into one giant EPR" (S8-RN-1)

Working with patient data that is dispersed even when it was produced in the same site, according to interviewees, is very challenging. This is mainly because it means that a lot of work and investment of resources is required before this data can be reused:

> So all the different teams have their own systems, which don't talk to each other. I think a lot of secondary care data is still handwritten which is a major challenge. So you have to go through the steps of someone having to digitise it to be able to use that kind of data and I haven't heard yet of there being any easy straightforward route to kind of use that data for research. (S6-AM-2)

This account illuminates some of the obstacles that thwart the possibility of bringing together patient data generated in the same site of data practice, ranging from the utilisation of different computer systems that cannot interoperate, to the utilisation of different data collection tools and methods. All this inevitably sets technical challenges that complicate the scenario for researchers when trying to access patient data from a single site of data production.

The challenges of working with fragmented datasets

Since patient data produced within the UK healthcare sector are fragmented and geographically dispersed, researchers wanting to access these data are required to approach a wide range of data providers, depending on the type of data they intend to use and the region where these data were produced. This finding confirms findings from previous studies focused on the flows of health data which have also pointed out that fragmentation is a key barrier for the reuse of patient data (Meystre et al., 2017; Weiskopf & Weng, 2013).

6.2.2 Access to rich but scattered pockets of patient data

Participants of this study, for example, have obtained access to patient data through a number of stakeholders, such as NHS Digital, SAIL Databank, Clinical Practice Research Datalink (CPRD), etc. While participants acknowledge that these sources give them the possibility to access rich datasets, they only offered scattered pockets of patient data with limited scale:

> You got discrete pools of very very high quality data around the country so like SAIL in Wales is like a really interesting database, really valuable data. You got CPRD. So you got all these pockets of quite rich data but none of them reach the scale what you need. (S3-EN-1)

This can be a barrier for research because it poses restrictions to the level of analysis researchers are able to make.

6.2.3 Difficult access to patient data that is not routinely collected

Some patient data produced within the UK healthcare sector are not available through sources such as NHS Digital, CPRD or any of the other sources mentioned before, and they are not shared outside the site where they are produced, unless the teams interested in using those data contact the site directly to negotiate and obtain permission to use these data.

Participants felt that data that do not form part of datasets available at intermediary sites of data practice such as Clinical Practice Research Datalink are particularly difficult to obtain. This issue was experienced by researchers on different data journeys (e.g. DJs 1 and 5). Researchers would need to do many additional technical arrangements, invest time in negotiations with information governance staff at sites of data production, and develop and implement measures to handle data according to the particular requirements established by the site of data production involved. This scenario is perceived as very challenging, particularly by early career and doctoral researchers with less experience and connections, because they would have to undertake a number of complex and time-consuming tasks in order to be able to access and reuse these datasets.

> I think if you want to use it, [patient data that have not been shared before] you would have to do a lot of work, you would have to spend a lot of time to try and work with hospitals. You would have to do that yourself as a researcher. Is not like you can get to CPRD and that already exist. You have to think about all the governance issues as well and anonymising the data and all that kind of thing. (S6-NM-1)

In summary, the convoluted data journeys within the healthcare sector set barriers that restrict the flow of this type of data to academic institutions to conduct research. This was particularly the case for researchers on the project *Long-term outcomes of urinary tract infection in childhood (LUCI)* that were dealing with data from different nations and researchers on the project *Using technology and data to improve the diagnosis and treatment of stroke*, who negotiated access to patient data directly with staff at health care provider organisations. Patient data produced within the healthcare sector are recorded, processed and stored in diverse ways according to governance frameworks and standards that are specific to their national context. In addition to the regional differences, there are also some discrepancies in the way the NHS providers manage data as primary care, secondary care, and tertiary care providers do not have a uniform way of managing patient data. This, according to participants of this study, has an

adverse effect on data flows, as it hinders the flow of patient data to academic institutions and can obstruct the advancement of research.

6.2.4 The development of Data Safe Havens

Another significant data infrastructure issue was around the functioning of data safe havens. As background to participant perspectives on this it may be useful to explain that Data Safe Havens (DSHs) are socio-technical solutions used across the healthcare sector that provide a suitable environment to enable storing, accessing and linking data generated in the healthcare sector securely and effectively within and beyond the NHS (Burton et al., 2015; The Academy of Medical Sciences, 2014). DSHs provide a way to maintain confidentiality and protect the patient's right to privacy (The Academy of Medical Sciences, 2014).

The first time the term 'Data Safe Haven' was formally used within the National Health Service (NHS) was at the beginning of the 1990s; the term referred to a physical place and the set of tasks and policies concerning the handling of patient data in a secure and confidential way. Over time, this meaning has evolved and is now widely used in the UK and internationally (Burton et al., 2015); however to date there is not a universally accepted definition of what a Data Safe Haven is; mainly because there is an ample variety of systems in operation. The NHS defines a DSH as the arrangements of administrative nature to ensure the safe and secure transfer of personally identifiable data across different sites (NHS Cambridgeshire and Peterborough Clinical Commissioning Group, 2017).

Independently of its origins and initial meaning, an entity that is referred to as a Data Safe Haven must enable the storing and releasing of data 'faithfully and effectively' (Burton et al., 2015). Besides that, a DSH must allow the storage and sharing of data in a way that might be perceived as trustworthy and secure by data generators, data users, research ethics committees, etc. (Burton et al., 2015).

The need for the utilisation of safe havens within the healthcare sector (NHS) is dictated by a number of Acts and guidance, including: 1) The Data Protection act 1998 (GOV.UK, 1998), 2) the NHS Digital Code of Practice on Confidential Information (NHS Digital, 2014), and 3) the NHS Information Governance Toolkit (Department of Health, UK, 2010).

The utilisation of data safe havens in the UK healthcare sector is considered important; this is because it has been recognised that the challenges in protecting personal healthcare records have significantly increased in the age of big data:

> Data is not necessary truly anonymous, because there are techniques and methods that can be used to re-identify people. So this is actually one of the key points is that individual record level data is never truly anonymised. What we try to do at the individual record level is to minimise the risk of re-identification and that is why you see thing like data safe havens…they all exist like that to put data in a safe managed place, because the fact that risk of reidentification is not zero. (S1-YR-1)

Data Safe Havens enable researchers in the academic sector to have greater access to data initially generated within the healthcare sector, and at the same time provide physical, administrative and technical controls (e.g. key/password management procedures, firewalls and audit logging) that act as barriers or frictions for the movement of data.

Slowness of systems

In this context, participants across all data journeys explored see Data Safe Havens as necessary means for storing and accessing data in a safe environment. The majority of interviewees agreed that it is paramount to have a safe environment for storing and accessing patient data for secondary uses. However, in their opinion Data Safe Havens do not operate in an efficient way, as they tend to set unnecessary and unjustified technical barriers for the utilisation of data produced in the healthcare sector:

> I think that the security around the data is appropriate. I think the problem is more about why the Safe Havens are rubbish I do think is appropriate to have our data really secured. (S3-EN-1)

Some data practitioners and researchers at sites of data reuse who were interviewed perceive as a major drawback the fact that they cannot work with data on the computers they use in their day-to-day work.

> Now that all happen within a safe haven environment, so it's quite slow…and it is so shit! It is slow; it crashes all the time. (S4-AS-1)

The utilisation of Data Safe Havens has been useful to provide safe means to share data but also constitutes a friction for the journeys of patient data as its use significantly slows down the data analysis process. The problem is not only that the operation of DSHs is extremely slow, but also that the system regularly stops working.

Physical barriers

Data in Safe Havens can only be accessed through secure access points, which are dedicated computers in an area that is physically secure where the utilisation or connection of external devices are not allowed and do not connect to the internet (Information Services Division, NHS National Services Scotland [ISD], n.d.). In general terms DSHs cannot be accessed remotely; however in some cases researchers are allowed remote access to the data via VPN (Royal

Academy of Engineering, n.d.) but this is rare. One participant explained how he accessed data from a dedicated secure room:

> So I work in a secure room when I work with patient data, there is a secure room, so I work there quite often on my own… I never leave anything unlocked, never leave room unlocked, etcetera, etcetera. (S4-NS-3)

Technical challenges

Given that Data Safe Havens do not connect to the internet, researchers experience difficulties because part of their work involves the utilisation of a number of programming languages for analysing data that require internet connection to operate in the most efficient way. As a researcher explains:

> So you know how R the programing language works? So a lot of it requires connection to the internet for reading packages and also you can directly download it from CRAN and it's constantly connected with the Internet. But if it's on a safe haven,…you want to read in a package, you got to …like do a really complex file transfer because, again…it doesn't connect to the Internet. (S3-EN-1)

Technical challenges to run programming languages within Data Safe Havens constitute an important barrier. Since researchers do not have the possibility of running programming languages in an optimal environment, they are required to carry out additional tasks that are hugely complex and time consuming to be able to analyse data.

The issues described here were experienced by researchers across all data journeys. The slowness of the system, physical barriers and technical challenges have been experienced by all researchers, from those working on small-scale projects to well-established teams with more resources.

Governance challenges

The constant crashes, the slowness in operation of data safe havens and the restriction to connect to internet are, according to researchers, concerning technical issues that significantly reduce the efficiency of Data Safe Havens and consequently slow down the movement of data between database and analyst. One of the governance issues experienced is that they are not allowed to export the results of the data analysis conducted in a DSH. For example, an interviewee commented:

> Another example is if you want to export your results from the safe haven to like PowerPoint. You have to literally email someone, to say can you please extract this from the safe haven? (S3-EN-1)

Besides the technical difficulties that researchers and data practitioners experience when working within data safe havens, participants commented that they also have to deal with challenges associated with the governance of Data Safe Havens. The existence of these issues

around data safe havens has generated a sense of frustration around users of health data. While participants recognise the importance and need of using this socio-technical solution to protect patient data, they believe that having to use Data Safe Havens significantly slows down the process of sharing and processing data for secondary uses. In their view, more than an aid to secure sharing of patient data, data safe havens act as a fundamental challenge for data to move.

6.2.5 Data quality

The NHS has indicated that as an organisation it values high data quality (NHS England, n.d.). According to NHS documentation and information provided on their website, they have adopted mechanisms to monitor the quality of data they produce and ensure their consistency, accuracy, timeliness, validity and completeness (Data Services for Commissioners (DSfC), NHS England, 2016; NHS Digital, 2018e). These mechanisms have been adopted specifically with the aim of providing high quality data driven by a need to deliver the right care to patients (primary use of data), monitor the performance of the organisation and comply with regulatory requirements (secondary uses of data).

Despite these ambitions, researchers and data practitioners interviewed as part of this research have expressed that it is not rare to come across data quality issues in data generated in the UK healthcare sector. As acknowledged in the literature review of this thesis, the fact that quality issues are commonly found in healthcare datasets has been also pointed out in previous studies within the health literature. Examples include the work of Meyste et al (2017) and Schlegel et al. (2017); both of these studies pointed out that users and reusers of data have often perceived data quality in healthcare records as insufficient. Such issues generate frictions to access, manipulate and analyse data.

Perceptions of quality

The majority of participants across all data journeys explored reported that most of the time they are able to access 'moderate quality' or 'low quality' data, as data are rarely 'high quality':

> You can get really quite, you can [get] reasonable data, moderate quality data, so you can never get high quality data. (S7-GN-2)

> So I don't know…for me it feels like we get poor quality data. (S6-AM-2)

Some participants who have worked with data generated at different sites have also pointed out that the level of quality varies depending on the type of site of data practice where it is produced:

> HES [Hospital Episode Statistics] data by and large is more accurate…in Primary care it doesn't work like that. (S6-HM-3)

Before going into detail about the data quality issues reported by participants, it is worth reminding that data quality is open to interpretation. The quality of data is evaluated differently by different stakeholders and could be defined by two interrelated factors: 1) to what extent it meets the expectations of individuals that use the data and 2) to what extent it represents the concepts, objects and events it is created to represent. Thus, it is not odd that researchers and data practitioners at sites of data reuse have a particular perception in regard to the quality of data produced in the healthcare sector that might be different to the views of healthcare professionals. Healthcare professionals are involved in the process of generating data and they use data for providing the right care to patients, whereas individuals at sites of data reuse are only users of data for secondary purposes.

Data gaps

Data gaps in patient health records are common across the NHS. The NHS itself has acknowledged this, and the organisation has adopted a number of strategies to address these gaps as they are considered damaging for the healthcare sector. Gaps in datasets can be different in nature and therefore, have different levels of impact in the work of researchers and data practitioners.

They can for example complicate and slow down the analysis of datasets, or in more serious cases they can inhibit the use of datasets as according to data researchers and data practitioners, missing data can be interpreted as a clear indication that a particular dataset is not suitable for doing analysis, or on other occasions they can lead researchers to question the validity of data.

As a participant explains:

> Yeah, yeah, there are problems with the data, so yes there is a lot of missing data for certain things, for example employment status, is not that well recorded probably about 50% is missing, even something like ethnicity, so you know which should be. ..so it should be recorded 100% full patients occasionally there is missing. (S5-IN-1)

Data gaps can be found in datasets produced at different sites of data production; however, these gaps can vary in nature depending on the site in which they were produced (e.g. GP practices, hospitals, clinics, etc.).

The fact that data gaps in health datasets generate frictions in the movement of patient data is a common finding identified in the literature exploring health data flows. As pointed out by a number of studies across the world, electronic health records are often incomplete (Kim et al., 2019; Thiru et al., 2003; Weiskopf & Weng, 2013). According to the majority of participants gaps in datasets is one of the key data quality issues that they have observed in their work with

patient. This confirms one of the key findings of Weiskopf & Weng (2013), which is that completeness is one of the most important dimensions taken into account when assessing the quality of data.

Datasets are not harmonised

According to participants across all data journeys, a major barrier for working with data from the healthcare sector is the lack of harmonisation in datasets, which generates inconsistencies. These inconsistencies are manifested in two different ways: 1) in the electronic record of a single patient where two or more contradictory data stories about the health of an individual are represented; 2) in the lack of consistency in the usage of classifications and coding schemes that are used across the healthcare sector.

- **Inconsistencies in the electronic record of a single patient.**

According to interviewees it is common to encounter inconsistent data in electronic healthcare records of patients. For instance, one participant commented:

> I spent time looking at smoking data and that can be recorded in various places in patient records and you can see examples. In one field one day a patient may had written that they are current smoker or a former smoker or on the same or never smokers, and then in a different field it may say how many cigarettes do you smoke a day? And it might say more than zero, so yes consistency is an issue. (S6-AM-2)

The key problem highlighted in this quote is that two different statements about the smoking practices of a patient are reported in the same electronic healthcare record. This is mainly problematic because it is impossible for those two statements to be true. However, for people at sites of data reuse, who are detached from the site of data production, it is very difficult to determine the validity and veracity of these data.

- **Clinical codes**

It is important to pay attention to data coding practices when studying the factors that influence the movement of data because these practices have a significant impact on the journey of data. Every decision made at the stage of coding data will affect future uses of the data recorded (e.g. the way data is coded will ease or complicate the retrieval of data).

Within the NHS, two clinical classification standards are used: OPCS-4 and ICD-10. These standards are focused on epidemiological and statistical analysis and the use of both of them is mandated at national level. OPCS-4 is used to classify interventions and surgical procedures (NHS Digital, 2019), whereas ICD-10 is a tool for classifying diseases and other conditions. Clinical coders at sites of data production are responsible for the classification of data derived

from patient records. Both standards are utilised to "support operational and strategic planning, resource utilisation, performance management, reimbursement, research and epidemiology" (NHS Digital, n.d.-c, para. 2).

According to interviewees, there is a lack of consistency in the use of coding and classification schemes across the healthcare sector, and this sets challenges for them at the stage of data analysis for various reasons. Whereas two standards are used (OPCS-4 and ICD-10), interviewees only reported experiencing issues with data coded with ICD-10.

As reported by participants, it is challenging to navigate data from the healthcare sector recorded in codes. In the first instance, because the majority of them do not have a clinical background; this however is not the only factor that makes it difficult for them to work with these data, particularly at the stage of data analysis. The main struggle for them is to decipher ambiguous meanings because symptoms and diagnoses, due to the design of the ICD system, can be coded in many different ways.

As one participant commented:

> So lot of medical data is recorded in read codes in kind of assigned code and if you don't know what you are looking for, as a non-clinician who does not do those codes that is quite tricky to navigate...So for example I am often looking for patients who have expressed some flu symptoms like cold...if you go into a GP, to SystemOne or Emis where we look to different records. There is about 100 different ways I have seen that entered, and you know, you know some would code it as one thing, someone as a another, someone don't code it at all and just write in text so it is not easy to work with that kind of data. (S6-NM-1)

Interviewees of this study believe that one of the reasons why there are so many discrepancies in the use of coding schemes is because clinicians are not completely familiar with all the existing codes.

As explained by one participant:

> They use different coding systems, and the coding systems are immense, so like ICD codes are, get between 200 and 300 thousand codes. GPs don't know their coding schemes. GPs are just selecting like typing a search term and then selecting from a list, they are not theme codes, they are not thinking about codes. (S7-GN-2)

Classification schemes, such as the ICD, aim to enable the systematic recording, analysis, interpretation of health data; however, this is not a simple endeavour. In the opinion of some participants, the categories of the ICD system need to be redesigned. However, it should be taken into account that no matter how good the scheme, it should be considered that there are never enough resources or trained personnel to manage a 100% accurate coding of data. Also, it is important to consider that patients could have several diagnoses, therefore different codes may be used when describing their health condition.

Reasons for quality issues

Participants expressed that they are aware that because they use data for a secondary purpose, they are likely to come upon a number of data quality issues that already sets a challenging scenario for them. This is a factor the majority of participants consider when setting their expectations in relation to data quality initially generated within the healthcare sector. For example, a researcher who worked on a large-scale project aimed at understanding age inequalities and inequities in the cancer pathway (Data Journey 5) expressed:

> All the data that we got in the repository, not all of them but most of them won't be designed for medical research. They would be designed for administration or for other purposes and so doing research with them is quite challenging. (S4-AS-1)

Participants commented that the limited amount of time that clinicians spend with each patient sets a challenging scenario to accurately record patient data. Clinicians are generally expected to make a diagnosis, give recommendations to the patient and record data in a system in a short consultation:

> The amount of time that clinicians have in clinical practice is limited, they simply don't have the time available to be able to robustly record in a very systematic way all of the data that is potentially able to be entered into the system, so yes there are time constraints. (S7-GN-2)

The data collection of some patient data is not mandated; given this situation, healthcare professionals and clinicians need to prioritise some tasks over others. As perceived by researchers and data practitioners across all data journeys, one key factor that influences the way data is recorded by healthcare professionals is the need for delivering the right care to patients:

> I would want the best quality data as possible, I want everything filled exactly as it should be, but I would imagine that someone working in a GP, that is not the priority. The priority for them is being able to treat the patient so they would want to collect only the information that would allow them to do that in whatever way. (S6-AM-2)

> Entering data is a very laborious process for them [the clinicians], it takes a lot of time. When they are with a patient they want to focus on a patient, they don't want to be writing things on a system. (S8-RA-2)

Some healthcare providers in England are part of the Payment by Results scheme. In this scheme, providers are paid for each patient they treat and patient data is recorded into a database managed at national level, and then the payment is calculated.

According to some participants, another factor that can influence the way data is recorded by healthcare professionals is the financial incentive that can be obtained if certain data is accurately recorded:

> In the UK, hospitals are paid based on how the data is coded, so HES, Hospital Episode Statistics, by and large be more accurate… so because hospitals get more money, hospital data is often more accurate because of that, because that's the way hospitals get paid. (S3-EN-1)

Participants believed that where a financial incentive is involved, codes are very accurate, whereas in settings where there are no financial incentives, data tends to be more inaccurate and data gaps are much more common.

For the majority of interviewees it is understandable that clinicians and other health professionals do not place the need to accurately record data as their number one priority:

> They are collecting the data for their own records you know, the patient in front of them is their primary concern… they are not necessarily thinking about the secondary uses of datasets (S6-HM-3)

> Their main focus quite rightly is caring for the patients and so, sharing data with a university is probably not high on their list of priorities. (S1-NR-3)

Interviewees expressed that they understand that the primary concern of healthcare professionals is the patient, and for this reason healthcare practitioners are not necessarily thinking about the secondary uses of data and making sure that they are coding everything nicely and always correctly.

Dealing with data quality issues

Since researchers and data practitioners cannot influence the data generation stage, they do not expect 'perfect data'. According to their testimonies, for them it is sufficient to be reassured that the level of quality would allow them to conduct the data analysis they intend to and to justify their conclusions. Dealing with data quality issues is, as perceived by interviewees, just part of their work as data is not always going to be perfect:

> So, the quality, that is a big a question for this kind of data. A lot of it comes down to accepting the quality is not always going to be perfect, and there is a lot of that. In writing up the research we have to acknowledge that there are these limitations. We are very much dependent on what data we got, because of course we can't influence that at the state of data collection. In terms of defining particular diagnosis and that kind of things, we tend to spend a lot of time thinking about how could you possibly do that with these data? (S6-AM-2)

Data quality issues found in data from the healthcare sector make it necessary for researchers and data practitioners to spend a lot of time conducting a series of checks to evaluate the datasets, cleaning data, creating descriptions of datasets, and manipulating the data 'to make them usable' prior to conducting any sort of analysis of data. Researchers and data practitioners interviewed consider that the main quality issues in data generated in the healthcare sector are that it is common to find data gaps and that datasets are not harmonised. Due to these issues they are required to spend a long time cleaning and making sense of data, which slows down the circulation of data.

As it can be observed, university-based researchers tend to perceive that 'patient datasets are not perfect', because they have gaps and contain errors. This suggests that they are aware that datasets generated in the healthcare sector are not infallible representations of the health of people, or in the words of Martin-Sanchez, one cannot expect that "such datasets offer a full or accurate clinical picture, let alone a full description of the health population under examination" (Martin-Sanchez et al., 2017, p. 30). They also recognise that it requires a lot of time and effort to make sense of data, that long hours of work cleaning data are necessary before they can make use of it, and that in some cases, even after cleaning datasets they might not be able to use them if significant data quality issues are identified. Nevertheless, as discussed in Chapter 5, when university-based researchers talk about the potential of patient data to 'save lives', a contradiction emerges: all this "work with data" of which they talk becomes blurry or less perceivable as these issues are rarely mentioned.

6.2.6 Metadata quality

Metadata are the structured information that describe data and facilitate their use and management data (Fabreau et al., 2018). For researchers and data practitioners at sites of data reuse it is important to have access to quality metadata in order to be able to identify the characteristics of the data that are available and then determine if those data are suitable for use in specific research projects.

Metadata are limited and lack in quality

As reported by participants, the metadata they have access to are limited and lack in quality. A number of participants across different data journeys commented that deciphering the characteristics of data and the way they are recorded is not straightforward; rather, it is a process with number of obstacles. Data practitioners and researchers commented that on occasions they get access to data dictionaries; however, these tools are not useful enough, as they do not provide accurate descriptions of data or sufficient information to fully grasp how datasets are integrated.

As a participant commented:

> You see a data dictionary and you'll see some documentation, look at it and you'll get some understanding of what it means but it is not until you start using those data but you realise what it really means and how it is really structured and what the errors are. (S4-NS-3)

Making sense of metadata

Participants pointed out that due to the lack of quality data, making sense of data involves a long learning process, as it is only when data practitioners and researchers embark on the data analysis journey that they begin to understand the composition of datasets, their errors and shortcomings. Participants who have been working with datasets with similar characteristics for longer periods find it easier to make sense of data than participants who are new in the field.

In this respect, a participant with several years of experience working with patient data who was part of the *Understanding age inequalities and inequities in the cancer pathway* team (Data Journey 5) explains:

> For example, in some of the health datasets for dates of admission or discharge from the hospital rather than being left blank, they stamped the 1st of January 1804, 1801 and there's nothing in the documentation telling you about that. Over time, when you try to work out the time that someone spent in hospital and you and you're getting some random days you go: "ok what does this mean?" And then you realise what are those errors, so any data that you get you will read it but not until you use it and then reuse it again you realise oh yeah learning, is a learning process. (S4-NS-3)

A number of participants working on large-scale projects or well-established research groups (e.g. Data Journey 1, *Using technology and data to improve the diagnosis and treatment of stroke,* Data Journey 2, *Building Rapid Interventions to Reduce Antibiotic Prescription* and Data Journey 5, *Understanding age inequalities and inequities in the cancer pathway*) commented that sometimes they receive guidance from clinical people to help them make sense of datasets; this has been useful for researchers to overcome the friction caused by the lack of quality metadata; however, not all teams of researchers have contact with clinicians who can orientate them. Having access to additional support from clinicians or healthcare professionals it is an advantage that often only those research groups with access to greater resources enjoy.

Dealing with metadata quality issues

In order to overcome the challenges generated by the lack of quality metadata and lubricate this friction, some groups of researchers have undertaken metadata projects. These projects are aimed at developing detailed descriptions about the data that are available at different sites of data production, not only to be able to use them in their own research projects but also to facilitate the access and use of these resources for other researchers at different sites of data reuse. A researcher from the *Building Rapid Interventions to Reduce Antibiotic Prescription* team (Data Journey 2) talks about one of these "metadata projects":

> I've got a colleague and he is trying to build a metadata catalogue to describe these data. To be able to say "Look this one Trust that's got an electronic wound care sheet that the nurses fill in when they are in their appointment is gathering these data" or "Look these data looks a bit like this… Hopefully it will help to illuminate to other researchers what is available. (S1-YR-1)

These efforts may help to overcome this friction to some extent; however, they are not definitive solutions because each data flow is unique and there are datasets that have never been shared before, therefore knowledge about these data has not yet been produced.

While data quality issues generate significant frictions for the movement of data, these interact with the expectations and assumptions that researchers hold about data. Given that university-based researchers are interested in using patient data because they hold the belief that these type of data have great potential and that conducting research with them is the most innovative way of exploring health issues, they are willing to enage in efforts to foster the movement of data towards universities. Evidence of this is the involvement of university-based researchers in initiatives aimed at improving the quality of data might help to reduce data frictions because if the quality of data is improved, researchers would spend less time cleaning and making sense of data. Nevertheles, it is worth pointing out that working on these metadata projects involve investing time and financial resources, therefore it is more likely that well-establishead research teams get involved in efforts of this nature.

6.3 Regulatory Frameworks

Regulations are an institutionalisation of norms, values and power relations. Regulatory frameworks that govern data flows in the UK healthcare sector aim to ensure the right balance between the protection of patients' confidentiality and the use and sharing of their personal data.

While in many cases the intention of regulations is to generate friction for purposes such as privacy protection, sometimes university-based researchers perceive them as too restrictive. They also felt that regulations sometimes generate unintended friction due to lack of clarity.

6.3.1 Unclear regulations

When asked about their views in relation to regulatory frameworks, participants expressed contrasting views. The majority of participants believed that the regulations controlling the flows and uses of data produced in the healthcare sector are excessively restrictive:

> I think probably some changes are needed, sometimes regulations are too restrictive (S7-GM-1)

In contrast, some other participants felt that the regulations are not too restrictive; however they pointed out that the problem with regulations is that they lack in clarity:

> I think they [guidelines and regulations] are not very clear. There is a lot of confusion about what needs to be done. I think that kind of put some barriers somehow… I know some people fairly enough they err on the side of caution. (S6-AM-2)

According to some participants, the fact that legislation is not always very clear and straightforward opens the possibility to interpret the legislation in two or more different ways and this generates confusion.

> What is challenging is that getting the permissions to do it, getting the IG and different people interpreting the law in different ways. (S4-AS-1)

This can mean that for example, researchers that request access to patient data with similar characteristics and under similar conditions from different sites of data production can have completely different outcomes depending on how each site of data production interprets the regulatory frameworks; one organisation may approve the data-sharing request, whereas the other may deny it. Participants who talked about unclear regulations as barriers for the movement of data towards universities were the researchers working in large research groups who have been involved in negotiations with different organisations for a single project, such as researchers working on data journeys 2, 3 and 5. For example, one participant who worked in the *Building Rapid Interventions to Reduce Antibiotic Prescription* project (Data Journey 2), commented:

> Then is the lack of clarity, some organisations are interpreting the rules differently to others. So you can go to these different organisations for the same approval and someone says yes, and someone says no. (S1-NR-3)

Interviewees tended to believe that due to the confusion that can be provoked by the lack of clarity in legislation and not having a clear path to follow, people prefer to be especially careful and not share data rather than taking a risk of making a mistake. Due to the lack of clarity, people do not have another way to find answers or solve their data sharing dilemmas, but go through a trial-and-error process:

> The rules are not black and white rules of what do you do…so they are interpreted in a variety of different ways and sadly until they get tested in court people start to realise what you can and can't do. (S4-SS-4)

This finding is similar to what Vikstrom and colleagues (2019) reported in a study conducted to examine barriers to the reuse of patient data for research purposes in Finland and Sweden. According to them, the lack of clarity in the guidance provided to request data for secondary uses can block or slow down the circulation of patient data; this is because instructions provided by data producers to support data requests are too generic, and therefore cause "interpretation issues among potential users".

A number of participants across different data journeys expressed the belief that some people in healthcare settings who do not want to share data have deliberately taken advantage of the lack of clarity in regulations and used legal loopholes as an opportunity to make excuses and

back their decisions to not share patient data. For example, a senior researcher working in the project 'Understanding age inequalities and inequities in the cancer pathway' (Data Journey 5) observed that due to lack of clarity they have experienced challenges accessing certain pockets of data for their project:

> Yes, different interpretations of the law and people hide behind the law. First of all they say they couldn't possibly do that because they are protecting the data, and then you say OK well here is my ethics, and it's been agreed that this is a good idea, and this how we are gonna make sure it's secure. And then they say oh…no no we can't do that…and then they come up with other excuses and I do feel often is just that they don't want to share. People like to own data if it's theirs they got a monopoly but is not their data is the patient data. (S4-AS-1)

This quotation suggests that the lack of clarity in legislation is perceived to not only pose a problem because some people in healthcare organisations struggle to understand the rules, but also because some use the lack of clarity to their advantage to create arguments that justify their decisions to not share data. In relation to this, it is also important to consider the existence of some tensions between law and ethics; something can be ethical even if illegal and vice versa.

6.3.2 Complex access processes

A major friction for data to flow from the healthcare sector to academic institutions, according to participants across all data journeys, except Data Journey 4, *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders*, was the complexity of navigating the varied access processes that exist to obtain permission to use patient data for secondary purposes. This difference, as explored in more detail below, could be because this was the only case in which no intermediaries were involved to grant access to the patient data used. Participants reported that they found it very challenging to keep abreast of the different set of rules and guidelines under which each data provider operates. They felt that having to go through several data application processes, which are different in nature with a number of data providers, is time consuming and overwhelming.

The governance to access patient data produced within the healthcare sector is not uniform across the four nations that are part of the United Kingdom and across different data providers.

Interviewees agreed that there is a need to establish access processes that work under a common set of guidelines in order to overcome the unnecessary complications that are characteristic of data access processes For example, a participant who was part of the team working on the project *Building Rapid Interventions to Reduce Antibiotic Prescription* (Data Journey 2) commented:

> There is inconsistency there in the different data providers, that has created a bit of a headache for me in terms of approving, we need to move towards a common way of working. (S1-YR-1)

The following quote from a participant working on the project *Long-term outcomes of urinary tract infection in childhood (LUCI)* exemplifies the impact that data access issues can make in a research project. Here one participant explained that one of the data journeys of one of the projects in which they were working was significantly slowed down because they had problems accessing data from one data provider:

> So you just go to go the same process with each data provider. So that is challenging, just navigating keeping on top of everything… where we were throughout, did incur a lot of delays with the project so we are actually over a year late reporting data simply because of data access issues, but we got them now. (S2-FA-1)

The majority of participants across all data journeys believe that obtaining access to patient data is complicated, independently of the route they take:

> Is hard work to get the data, you know is painful to go through all these access process. (S1-YR-1)

Almost all participants expressed that they have struggled to access patient data, however, it can be said that the most challenging route is the one that involved direct negotiations with sites of data production, as referred by participants in three of the five data journeys explored (DJs 1, 3, and 5). A senior researcher leading the research project 'Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders' (Data Journey 4), expressed a contrasting view. For this participant, accessing patient data for their project has been always very smooth, easy and straightforward. They commented that they had never experienced any complications in accessing patient data:

> So I think the system we have locally works very well. So I never had problems with restrictions which have affected the quality or the amount of time is taking to do my research. (S5-IN-1)

It can be suggested that this experience was influenced by the fact that researchers working on this project negotatiate access to patient data through a local committee because the site of data production and the site of data reuse were part of the same institution.

Data requests that are processed by a local committee, in contrast to other data application processes, do not require the intervention of intermediaries. That is to say, that it is not necessary to apply for National Ethical approval or obtain permissions to access patient data from external organisations such as NHS Digital or CPRD. Local committees are responsible for ensuring that research data applications are within ethical and legal guidelines.

Different rules apply for each nation

The lack of uniformity of the rules for accessing data across the four nations, according to the majority of interviewees across all data journeys, sets a complex scenario for researchers. For example, some specific types of patient data that cannot be shared with certain types of stakeholders in one nation can flow with less friction in another. As this participant explains:

> Scotland has different rules than England. So, like in Scotland you are not meant to share data with pharma companies or insurance companies and people like that, unless they are affiliated with an academic and I think that is arbitrary. (S7-GN-2)

The above quote suggests that in some regions, some forms of patient data with similar characteristics move with fewer constraints or less friction from sites of data production to sites of data reuse.

Different time frames and systems

Each data provider sets their own working time frames, and uses different systems to process data requests. In addition to this, the questions that data requesters are asked when they submit an application to obtain access patient data are not always the same. Some interviewees who have had to negotiate access with different data providers for the same project expressed similar views. For example a researcher working in the research project *Long-term outcomes of urinary tract infection in childhood* (Data Journey 3) for which data was requested from English and Scotish data providers commented:

> Different data providers want different information… they all ask very different things, the application process is different, time frames are different. (S2-FA-1)

Similarly, a researcher working on the project *Understanding age inequalities and inequities in the cancer pathway* (DJ 5) which used data from six different types of NHS service providers expressed:

> Different data owners, everyone's got different systems…They have their own data application process. (S4-AS-1)

The fact that data providers work under different time frames and process data applications in different ways is problematic for researchers working with data coming from different sites of data production. Mainly due to the fragmentation in datasets, they are often required to request data from more than one data provider, thus they need to become familiar with the data application processes of several data providers and try and keep up with the time frames set by each of them.

Different conditions on data sharing agreements

In addition to the fact that the length of time of application processes and systems are different among data providers, participants also have to deal with the issue that data sharing agreements are not uniform; rather they can be drafted in different ways and can take a variety of forms. Each data provider sets their standards for handing data, meaning that researchers are required to adopt different data processing or storage procedures depending on the data provider involved. This was particularly challenging for researchers working with data from various organisations. A member of the *Long-term outcomes of urinary tract infection in childhood (LUCI)* team explained:

> We hold data from quite a few different organisations, they all have data sharing agreements and so they all require different things, so they might require the data to be held differently or to be stored differently. (S2-FA-1)

The following quote sheds light on some of the differences that exist between data application processes within the UK. This extract suggests that the processes to access patient data generated within the healthcare sector are all very different from one another:

> Each one [each application process] is slightly different. So NHS Digital they have an online application process and then the Scottish system, there are a lot of different little datasets that you can get on one application in Scotland but it is no always straightforward, sometimes it is slightly more complicated. Then in Wales, the data is in Swansea...that comes to GP data and many other datasets. So the application process can be quite time consuming. (S7-GN-2)

Each data access process requires researchers to undertake a number of tasks that can be difficult to deal with. These processes have different levels of difficulty and researchers find it challenging; they struggle to navigate them and perceive them as time consuming.

To summarise, regulatory frameworks at times create unintended frictions that can slow down or block the circulation of data from the healthcare sector to academic institutions. The lack of clarity is one of the factors that has contributed to this, as data providers might have difficulties in judging whether they should share the requested data or not, and ultimately decide to deny access. The lack of consistency among processes to obtain access to patient data can also create barriers for data to move. Researchers are often required to develop an understanding of different regulations before seeking access to patient data, which can be time consuming, particulary those working with patient data coming from different sites of data production (e.g. DJs 2, 3 and 5). In contrast, this does not represent a significant barrier for researchers working with patient data from a single site of data production as was the case of the project *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders* (Data Journey 4).

Some research teams are impacted in the sense that their access to data can be slower than expected, whereas other teams with less resources might give up seeking access to data if they struggle to make sense of regulations. In addition to this, since data providers work under different time frames and conditions, researchers are required to conduct different procedures and make different arrangements to meet the requirements of data providers, which again could slow down the circulation of patient data to universities.

6.4 Sociocultural factors

This section explores the sociocultural factors that create barriers for patient data created within the healthcare sector to move to sites of data reuse in academic institutions. As explained earlier in this work, in the context of this study, sociocultural factors are understood as 1) motivations for doing something or engaging in a task or activity, 2) social norms 3) beliefs and expectations, and 3) desires, affects or wishes, all of them shared by a group of people or a group of professionals

6.4.1 Different priorities across sites of data practice

As explained in a previous section of this chapter, one of the routes to access patient data to be reused for research in the academic contexts is through a direct negotiation with people at the site where those specific data were produced. The majority of interviewees that participated in this research have been involved in projects where they were required to negotiate access to patient data directly with information governance staff at sites of data production.

Interviewees commented that the negotiations with information governance staff are difficult and that obtaining access to patient data is very challenging. Several participants across all data journeys commented that one of the greatest frictions they have encountered to access patient data is the reluctance of healthcare professionals and information governance staff at sites of data production to share data for secondary purposes with stakeholders outside the healthcare sector. Participants believed that this is due to a number of factors: 1) sharing data beyond the healthcare sector is not a matter of priority for stakeholders at data production sites; 2) the existence of a deeply ingrained risk averse culture across the healthcare sector; 3) lack of familiarity with relevant regulations; and 4) and the fear of being criticised. While these are the perceptions expressed by university-based researchers, the literature exploring the views of healthcare professionals offered contrasting insights. Previous studies suggest that although the main priority of professionals working in the healthcare sector is to use patient data to provide

direct care to patients, they recognise the benefits and value of reusing patient data (Neves et al., 2019). Existing research also suggests that staff working within the healthcare sector do not always oppose sharing patient data beyond the healthcare sector; nevertheless, they are aware of the potential negative consequences that could emerge if the reuse of patient data is not appropriately conducted (Mouton et al., 2018; Neves et al., 2019). This will be explored in more detail later in this chapter.

6.4.2 The resistance of healthcare institutions to share patient data

According to participants across all the different data journeys expored, the negotiations with people in the healthcare sector are very difficult. One of the reasons for this, as perceived by participants of this research, is that staff at healthcare organisations do not see sharing data for secondary purposes as a priority. This, according to them, becomes evident in the fact that at times healthcare professionals oppose data sharing projects or avoid participating in them. This, in their perception, has made negotiations particularly difficult:

> There seems to be among information governance staff a lot of them saying no, there is a don't say yes culture, there is not a 'oh yes, this needs to happen'. (S1-YR-1)
>
> They don't want to share data, a lot of them, or don't see the point of it. (S7-GN-2)

Participants believed that whereas information governance staff do not tend to express that they are against data sharing openly, their resistance to this becomes evident by the way they respond to the requests made by organisations outside the healthcare sector who intend to use data for secondary purposes. As a result of this, the negotiations to obtain access to data becomes a push and pull dynamic between the parties involved (data production sites and data reuse sites). This negotiation process is perceived more as an exercise in which researchers' objective is to debunk the unfounded excuses of data providers.

While it is true that perhaps decision-makers in the healthcare setting might show reluctance to share data, it is also possible that resistance for sharing data is driven by genuine concerns that university-based researchers do not recognise because their ethical values and priority systems might be different.

The way in which negotiation processes between researchers and healthcare institutions unfold reveals the existence of a power dynamic between these two stakeholders. Healthcare institutions have power in the sense that they have control over the data. On the other hand, research groups have a technical power stemming from their expertise. It would seem that they

have the expectation and feel entitled to access data so that they can use their skills to conduct research under their own terms, and according to their own ethical guidelines.

6.4.3 Perception of a risk averse culture within healthcare institutions

One of the major frictions that was perceived to slow down or block the flow of patient data generated within the healthcare sector was the belief that a deep-rooted risk-averse culture is ingrained among information governance staff and healthcare professionals within the NHS and other key stakeholders:

> The clinical professionals, are very risk averse, those who hold the data are very risk averse, NHS Digital, Public Health England. (S4-SS-4)

University-based researchers across all data journeys expressed that the risk aversion significantly complicates the communication and negotiations with people at healthcare organisations. Findings of this study suggest that university-based researchers point out the existence of a risk-aversion culture, but without a real engagement or reflection with the reasons that might lead data providers to sometimes refuse sharing patient data. Contrary to what university-based researchers express, frictions created by the healthcare sector might be quite reasonable. Rather than a statement of fact, the risk aversion accusation could be a political movement to portray the opposition to data sharing in a bad light.

According to the majority of interviewees, people tend to be that scared that it is very complicated to engage in a conversation with them. One participant commented:

> They [healthcare professionals] are very extremely scared. So, essentially, is very difficult to interact with them. (S7-GN-2)

In the above quote, it is suggested that staff at healthcare organisations are driven by their emotions as the interviewee talks about them as being "very extremely scared". Whereas emotions are usually involved in any decision making, even if not consciously, it is very unlikely that staff act solely on their emotions.

Some participants felt that risk-averse people within the healthcare sector have difficulty taking into account the legitimate evidence presented by university-based research teams aimed at demonstrating they meet the requirements to ensure data they obtain is going to be protected and adequately handled. As one participant commented:

> There seems to be among information governance staff a lot of them saying no, there is a don't say yes culture, there is not a 'oh yes, this needs to happen'…the people who need to sign yes to a sharing agreement they are often really reluctant to say yes, even if you tick all the boxes and show how you go over and above the organisation itself in order to protect the data. (S1-YR-1)

The above quote suggests that perhaps some university-based researchers see the data requests application process as a tick-box exercise without engaging in or acknowledging the more substantive ethical issues.

This participant further shared an anecdote in which they explained the challenges experienced in obtaining access to patient data. In this account, they explained how a project that attempted to use pseudonymised patient data from three different sites of data production could not be realised because information governance people at one of the sites of data production involved declined to share patient data. Since this project required having access to the data produced at the three sites of data practice in order to be able to realise the project, even having the two other stakeholders on board, they could not conduct the project:

> There was one hospital that was very keen on it, but because the other big players didn't really buy into it, there was one particular, there was a big merge of hospitals and there was a pilot going on with one aspect of one hospital that merged. But they merged with a very risk averse trust who don't like to share patient level information with researchers even if it is pseudonymised, they are like 'No!' There was a nervousness around, there was a special board that would decide on whether data that had been integrated could be shared with researchers, and I don't know. (S1-YR-1)

In this quote, in a similar way to an earlier extract, the participant suggest that the responses of staff at healthcare organisations are solely driven by emotions. They do not seem to consider the fact that perhaps the reason for why the trust did not agree to share data was because pseudonymisation does not guarantee that individuals cannot be identified.

Several interviewees across all the different data journeys explored commented that despite their data requests clearly demonstrating a lawful intention for using patient data and that they are committed to following the guidelines established to handle data, their requests have been rejected due to the lack of familiarity of staff at healthcare organisations with regulations and their risk aversion. For example, a participant mentioned:

> Even when there is a legitimate need and there is a legal basis, there are people being very risk averse, not really understanding information governance legislation, not really understanding the duty to share. (S4-AS-1)

In the above quotation, the participant highlights the perception that healthcare professionals do not really understand their duty to share as if this duty establishes that healthcare professionals have the responsibility to share data for secondary purposes. However, in reality this "duty to share" (Department of Health, 2013a) refers to the responsibility that healthcare professionals have to share data to support the direct care of patients within the healthcare sector, rather than for secondary purposes.

Another participant who pointed out to the lack of understanding of laws among staff in the healthcare sector mentioned:

> A great misunderstanding around GDPR among the practices and it makes them much more hesitant to share data and be nervous about sharing data, so it has been a huge problem. (S7-GN-2)

Participants tended to highlight that their research project proposals were in line with all the legal requirements to handle patient data. However, as noted above the law lacks clarity sometimes. Therefore, it is worth raising the question of whether researchers can be sure that legal requirements were met, or if this was just an assumption made by them.

According to participants across all data journeys explored, it is common to see that when people in the healthcare sector need to decide whether to share or not share patient data without having a clear understanding of what is the right thing to do, they tend to stay on the safe side and do not share data to avoid making a wrong decision:

> A lot of people are very anxious, they are very cautious about doing anything wrong, so they prefer to not do anything, because you cannot do anything wrong if you don't do anything…so I suppose that is the complexity. (S4-LS-2)

Participants that have experienced challenges accessing patient data have made efforts to ease this friction. For example, one participant shared their experience in a project for which they needed to obtain access to patient data generated at a number of general practices. The team started the negotiation with the sites of data production; however information governance staff at the sites of data production showed a lot of resistance to sharing data because they were unsure whether the data request was law compliant or not. The team of researchers then approached the Information Commissioner's Office to ask them for their support by communicating in written form with GP practices and reassuring them that the data request made by this group of researchers was lawful and that they were not putting their organisation at risk by sharing this data. The ICO wrote a letter in support of the project. The fact that despite these efforts, some GP practices still did not agree to share data was disappointing for this interviewee, who commented:

> So the issue that we are having in particular is with GDPR, practices have to do a DPIA, and they are concerned they could get reported to the ICO and be fined. So the GDPR, but in fact, I mean the thing that kills me about this is essentially we went to the ICO, we spoke to the Information Commissioner herself and she wrote in support of our project linked GP data and we can provided them with a copy of this letter and they still won't agree. (S7-GN-2)

The quote above suggests that people within the healthcare sector refuse to change their stance in relation to data sharing even when researchers provide them with explanations for why it was lawful and safe to share patient data. This example draws attention to legal and ethical

tensions that can emerge. Even if something is lawful, some people still might have concerns about whether it is ethical and in the best interests of patients, who are the healthcare sector's primary responsibility.

As it can be seen, some university-based researchers perceive that one of the main reasons people working at healthcare organisations are reluctant to share patient data is because they are extremely risk averse and because they are not aware of the benefits of conducting research with patient data. University-based researchers seem to believe that having a legal justification for accessing patient data should help to smooth data flows. Nevertheless, as findings from previous research suggests, for practitioners in healthcare while it is essential to comply with legal frameworks, for them it is also important to ensure the ethical use of patient data (Neves et al., 2019).

6.4.4 Evidence provided by previous research exploring the views of staff in healthcare settings shows that healthcare professionals do not always oppose the reuse of patient data (Ford et al., 2020; Neves et al., 2019). Nevertheless, they are aware and have expressed concern regarding the potential negative consequences of sharing data beyond the healthcare sector. For example the potential re-identification of patients, the use of research results that could damage the already vulnerable groups (Ford et al., 2020; Mouton et al., 2018). **Fear of criticism within healthcare institutions**

Some participants commented that the fear of being criticised by the mass media or the public is another trigger for the reluctance of healthcare staff to share patient data beyond the healthcare sector. For example, one participant commented:

> You can get criticised either way so if you share you can get criticised, if you don't share you can get criticised. So I think sometimes they are more on the side of caution and not giving access because of the things that have been on the press stories so they are worried to being criticised. (S4-LS-2)

Similarly, another participant expressed:

> They are scared about how things may be portrayed and that [they will be] be criticised. (S4-YS-5)

Participants tended to express the belief that maintaining a good public image and protecting their reputation is a matter of great concern for people within the healthcare sector. In the two quotes presented above, staff from healthcare organisations are presented as fearful of public opinion. This demonstrates that university-based researchers have been inclined to frame the actions of staff at healthcare organisations as emotional responses.

6.4.5 The negative effect of Care.data

Several participants across all the different data journeys explored in this study commented that Care.data, the controversial initiative (now cancelled) which has the objective of sharing patient data generated in primary care with external organisations without informing patients what data would be shared and to whom, significantly damaged data flows in the UK healthcare sector.

According to participants, two of the major issues associated with the Care.data initiative were its lack of transparency and its failure in gaining public trust. This has had a negative impact on data flows from the healthcare sector through to sites of data reuse in the academic context. As one interviewee explained:

> It wasn't transparent, it wasn't communicating very well. It wasn't well thought through, there was no patient involvement at all. It was a good example of how not to do it, it was very untransparent [sic] if that is a word and it has massive effect. (S4-SS-4)

This quote suggests that university-based researchers perceived that data flows became complicated after the Care.data controversy.

Frozen data flows

Participants who were working on projects using data produced in the healthcare sector at the time when Care.data was under public scrutiny explained that they started experiencing the negative consequences of this initiative shortly after the intentions of Care.data came to the light, this was the case for example of researchers working on the project *Understanding age inequalities and inequities in the cancer pathway*. They attributed the complications they started having trying to access patient data and making data flow from the healthcare sector to the academic sector for conducting research to the Care.data scandal. One of the researchers from this research team commented:

> Data wasn't being used, researchers couldn't get access to data…for this reason of Care.data and lack of transparency. (S4-AS-1)

Similarly, another participant from this same group of researchers reported:

> About 4 years ago, there was a huge issue around lack of access to data for research, because of Care.data and the data flows to research stopped" (S4-SS-4)

These quotes suggest that these participants perceived that the access to patient data for research became more complicated as an immediate consequence of the Care.data initiative. From their perspective, negotiations with sites of data production became more difficult, and some projects were delayed, whereas others were completely stopped.

Number of opt-outs increased

Another consequence of the Care.data scandal was that after this, a large number of patients decided to stop allowing the NHS to share their data to be reused for research purposes outside the healthcare sector. After Care.data a lot of people decided to stop sharing their data because they were afraid their data was at risk of being misused. This significantly affected the flows of patient data because even after the cancellation of Care.data, those people who opted out remained in the list. As reported by one participant:

> When Care.data happen there were lots of people who wanted to opt out of it, GPs were giving this. .opt out. ..you know you were aware of those, and at some point 6 million of people said I don't want my data to be used outside of direct care. Even when Care.data was stopped, scrapped, those people remained on the list of opt outs. And those opt outs are now affecting data flows going out to NHS digital and indeed now Public Health England. (S4-SS-4)

According to the perceptions of university-based researchers and data practitioners across all data journeys, the controversial Care.data initiative has generated increased concern in relation to privacy risks and some patients decided to take some action to generate friction.

Academic researchers are still paying the price of Care.data impact

The aftermath of the Care.data scandal, in the opinion of participants, is still affecting flows of patient data from the healthcare sector to sites of data reuse. Participants across all data journeys felt that the public lost confidence in data sharing initiatives, even in those led by academic organisations, and that the problems in accessing patient data generated by Care.data have persisted over time. They have the feeling that even up to now, researchers are still struggling to restore the broken data flows. This perception is exemplified by the comment of a researcher working on *Understanding age inequalities and inequities in the cancer pathway* (Data Journey 5) who said:

> Then the Care.data scandal in which we lost all the public confidence in how data we use and the whole system just frozen, no data was going anyway, now we couldn't analyse anything which was tragic…so all fell to pieces and then gradually we've been trying to build it back up again (S4-AS-1)

Research shows that while healthcare professionals supported the Care.data initiative when it was announced, some of them later criticised and opposed this initiative (Ford et al., 2020). This happened mainly because they perceived that there was inappropriately managed, it was not clear who and under which conditions were going to have access to patient data, and little information about the scheme was provided to patients.

As per the views of the public, when transparency and privacy issues regarding the Care.data scheme were revealed, members of the public heavily criticized this initiative in social media

(Hays & Daker-White, 2015) and more than 1.5 million opted-out from their data for being shared beyond the healthcare sector. Nevertheless, while scandals can negatively affect levels of trust and confidence in the ability of the National Health Service to protect healthcare data, support for the reuse of patient data for research purposes if this is for the public benefit has remained high (Healthwatch England, 2018).

6.4.6 Bigger institutions, less frictions

The process of obtaining patient data to be reused in the healthcare sector is complex and involves several stages that can be time consuming, requiring investment of large amounts of money and negotiations with a number of stakeholders at several organisations. When reflecting on the challenges to access patient data, participants commented that they acknowledge that bigger organisations or prestigious research groups with access to more resources are in a better position to overcome the frictions to access patient data generated within the healthcare sector than smaller institutions or early career researchers.

A number of participants across all data journeys commented that processes to access patient data are too bureaucratic and expensive, and that researchers with limited resources are the most heavily affected by this fact. For example, a member of the *Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders* team expressed:

> They always disadvantage people with the least resources and is essentially that is just crazy bureaucracy. (S5-IN-1)

This participant later added that the access processes that are needed to obtain patient data alongside the large amounts of money that need to be paid can limit the progress of medical research:

> I think the various processes you have to go through, particularly I would say for smaller research groups or researchers who don't have much money or having to do things by themselves it can mean that people are put off doing any kind of research and I think it does hold back the progress of medical research. (S5-IN-1)

Interviewees who work for large and prestigious institutions or research groups with access to a large pool or resources or have close relations with people at senior positions at key organisations acknowledged that they are in a privileged place compared with smaller research groups or early career researchers and because of this, they can access larger amounts of patient data. As one participant commented:

> We have a person who is assigned to the project and I have his number so if I have any issue I can call him. The senior people within this project are in touch with the senior people in NHS Digital and they are quite supportive of us, so as I say I am in a privileged position, you know most other projects can't get that kind of support from very senior people that I do, we are getting more data than other people. (S7-GN-2)

This participant further added:

> I always remember if we cannot get the data must other projects won't be able to because we are considered one of the most important and robust projects in the UK. So I think generally have a much easier time than the average researcher cause is quite a bit of a political pressure that we be brought to there if we say NHS Digital won't give us data. (S7-GN-2)

As acknowledged by researchers working on bigger organisations or research groups, such as participants of data journeys 1, 2 and 5, while they do not have a frictionless access to patient data, they do have some benefits and better chances to obtain patient data generated in the healthcare sector to conduct research. This is because they can exert political pressure if they struggle to obtain patient data.

6.5 Chapter conclusion

This chapter presented the socio-material factors that have generated frictions for the movement of patient data towards university-based research projects or groups.

According to findings of this research, Data sharing infrastructures and management generate frictions for the movement of data. Whereas significant amounts of patient data are generated, these data are held by different organisations and managed in different ways. Patient data are managed according to governance frameworks and standards that are specific to their national context. In addition to the regional differences, there are also some discrepancies in the way the NHS providers manage data. This, according to participants of this study, hinders the flow of patient data to universities. Not having all patient data available in a single source can slow down or block the movement of data due to technical challenges that can emerge as a result of this. Another significant data infrastructure issue was around the functioning of Data Safe Havens. While Data Safe Havens have been useful to provide safe means to share data, they do not always operate in an efficient way and their use significantly slows down the data analysis process. One of the main issues is that often they can only be accessed in specific physical places where the utilisation or connection of external devices is not allowed and the connection to the Internet is restricted. These findings confirm previous research, a number of studies have reported that fragmentation of data sets and issues with infrastructures can significantly slow down or completely stop the movement of patient data (Meystre et al., 2017; Weiskopf & Weng, 2013).

The majority of participants reported quality issues in patient data. According to their accounts, data gaps are commonly found in these datasets. Additionally, there is a lack of consistency in the use of coding and classification schemes across the healthcare sector. For this reason they are required to spend a long time cleaning and making sense of data. This is also not a new finding. A number of previous studies have demonstrated that it is very common to data quality issues in the datasets produced in the healthcare sector and that this slows down and complicates the work of researchers, in other words, this can slow down or block the movement of patient data. As Schlegel and colleagues (2017) have pointed out, it is always necessary to invest significant amount of time preparing data before it can even be reused. It was observed that the expectations and assumptions that researchers hold about data interact with the issues to access and use data produced by metadata quality issues. As explained above, givent that university-based researchers believe that patient data are resources with great potential and that conducting research data with them is the most innovative way of exploring health issues, in order to overcome the challenges generated by the lack of quality metadata and lubricate this friction, some groups of researchers have engaged in projects aimed at improving the quality of datasets. They have contributed to producing detailed descriptions about the data that are available at different sites of data production to be able to use them in their own research projects and to facilitate the access and use of these resources for other researchers. Efforts for overcoming or lubricating frictions have been observed in other communities. For example as Bates and colleagues (2019) pointed out, practitioners working with climate data offered their voluntary work in order to cope with the challenges for using data posed by the infrastructure.

University-based researchers and data practitioners have the perception that a risk-averse culture within the healthcare sector generates barriers for the flow of patient data to universities. They would expect to be granted access to patient data on the grounds that their requirements are legally defensible. It would seem also that university-based researchers have the perception that healthcare professionals do not understand the benefits of sharing data for research purposes. As discussed in this chapter this claim of researchers is not validated by studies that have paid attention to the perceptions of healthcare professionals. Previous research shows that many practitioners in the healthcare sector are aware of the benefits of conducting research with patient data and are willing to support the reuse of patient data if adequate measures are in place (Ford et al., 2020; Neves et al., 2019).

Drawing from the findings of this chapter, it can be suggested that university-based researchers do not have a clear understanding of the motivations of practitioners at site of data production

for supporting or showing reluctance for sharing patient data. This could generate or exacerbate the already existing tension between these different cultures. The frustration resulting from the perception that staff at sites of data production most of the time are reluctant to share patient data interacts with the frustration around the complexity of regulatory frameworks, physical barriers, and technical challenges to create a sense of significant friction from a variety of social and material sources.

Chapter 7. The voice of patients and the public: a key strategy to smooth data frictions

According to participants, as addressed in detail in the "frictions" section of this work, after the Care.data scandal, it became very difficult to make patient data flow from the healthcare sector to external sites as the public expressed discontent with the idea of sharing data for purposes beyond the direct care of patients. As discussed in the previous chapter, this, according to the perceptions of university-based researchers was in part because people working at healthcare providing sites became very anxious, did not wanted to share data and were too worried, it was also because members of the public stop trusting external sites beyond the NHS with their data and therefore became less supportive of patient data sharing beyond healthcare settings.

A number of participants across all the different data journeys explored agreed that one of the main reasons that led to the failure of Care.data was that in this project neither the patient nor the public voice were taken into account. As told by some participants of this study, they have always considered it important to take into account patients and the public view; nevertheless, the opinions of patients and the public in relation to the uses of patient data became more salient after the negative consequences of Care.data, which started to emerge and become evident in 2014.

Some participants commented that after Care.data the academic community engaged in a deeper reflection about how they had integrated the views of patients and the public in their past research and started designing and implementing strategies and undertaking efforts to make sure they integrated the views of patients and the public in their projects. This remark came mainly from participants who are part of projects that regularly engage with groups of patients or members of the public (DJs 2, 3 and 5).

The importance of the public and patient involvement is featured in a number of the research projects' websites for some of the data journeys explored in this study (e.g. *Building Rapid Interventions to Reduce Antibiotic Prescription* (Data Journey 2) and *Understanding age inequalities and inequities in the cancer pathway* (Data Journey 5). The important role of patients and the public is also clearly evident in the research focus of health studies on data flows and reuse described in section 2.3.2 of the literature review of this thesis. The majority of interviewees across all the data journeys explored commented that their research projects receive some sort of input from patients, the public, or from both of these groups of stakeholders. Some of them have patient groups that interact with them on a monthly basis, whereas others have citizen juries and more active patient groups. For example, a participant a member of the *Long-term outcomes of urinary tract infection in childhood (LUCI)* team (Data Journey 3) explained, in relation to the involvement of the public, that their organisation had placed an important role on the voice of the public and the patients, thus project leads are always required to take into account the views of members of the public:

> The whole time I worked here, it has always been like you have to have members of the public on one of your panels, because half of the time we just make the assumptions we are like we will do this, how do we know that that would be acceptable, so it is definitely worth having a number of people with different perspectives... we are really interested in the public perspective side, acceptability definitely will always be an important aspect. (S2-FA-1)

The above quote shows that one of the reasons why researchers consider it important to know the views of the public is because there is an interest in understanding what uses of data are considered acceptable for them, and this, according to this participant, is better than making decisions based on assumptions.

The engagement with patients and members of the public has helped to drive data towards university-based research groups. First, the engagement with public and patient groups gives university-based research groups credibility and legitimacy in the eyes of funders, the public and data providers. In some instances where people at sites of data production have shown

reluctance to share data, the argument that researchers have the support of patients has been used to try to convince data providers to agree to share data.

7.1 Patient and public opinion to mitigate risk aversion

According to some participants, on several occasions requests for data are denied not because of the real risks associated with data sharing, or because their projects fail to meet the requirements to keep the data secure and safe, but because of the risk aversion of key decision makers at the site of data production. As commented by some participants, particularly those from well-established and recognised research groups that engage with groups of patients on a regular basis one way of contesting this risk aversion is by bringing forward the opinions of patients and the public. For example, a participant working on *Understanding age inequalities and inequities in the cancer pathway* commented that the opinions of patients can be useful to overcome frictions that some data journeys encounter.. In this sense it can be a factor that could ease the friction generated by the risk aversion acting as a lubrication factor.

> So for me, the patient voice around the data work is particularly critical, there's so much data, so much we could do but there is so much risk aversion, the only thing that is going to unblock it is the patient's voice that is the only way we will unblock the advances we need to, as simple as that. (S4-SS-4)

Another participant, from the same research group whose role involves liaising with key decision makers and participating in negotiations to obtain access to patient data, explained that at times when negotiations had become too complicated for her and decision makers resisted sharing data under the argument that doing it is risky or unsafe, she tends to bring in the views gathered in meetings with their "public-patient" group as her counterargument. According to this participant, the justifications provided by people at sites of data production are often just excuses. This participant commented:

> When someone says I cannot give you the data because is a risk, when people start arguing about what they can or they can't they always say "well patients don't want you to use it" or "you cannot have the data because I am protecting it" well we say here are patients, what do you think? and they can start to argue… I say well actually I asked my patient group what they think and they assess whether the mechanisms we put in place are secure, you know how they think. If they say, people have to put their justifications of what they are not gonna share the data or collaborate, then we can put annotations and say well I think the benefits are bigger than the risks of that and see what you think, so I try to keep step back of it, is quite a challenge. (S4-AS-1)

This participant further explained that when negotiations turn complicated, she takes an approach of bringing forward and drawing the attention of people at sites of data production to the voice of patients rather than giving justifications developed by their research team. She referred to the patient opinion as her "biggest weapon" to ease the friction generated by the reluctance of people at sites of data production to share data:

> Actually, my biggest weapon in this is the patients, and if I have to step back from the data, it doesn't become my data, it is the patients' data and they can be the ones that are betrayed about what happens. (S4-AS-1)

This quotation not only shows the way in which this researcher presents the voice of patients to people at sites of data production, but also shows that for some participants, not sharing data for research purposes is perceived as a betrayal to patients.

This suggests that perhaps researchers can sometimes overlook the fact that not all people have the same concerns. As the literature has demonstrated listening to some patients, small groups of patients is not equal to listening the concerns of patients with different backgrounds, situations in life and concerns (Understanding Patient Data, 2018).

Another participant shared a similar idea to the one presented below since according to him, to use data for research is a responsibility they have with the patients:

> I think there is a responsibility to the patients to use the data that has been collected. (S6-HM-3)

Using the support of patients as an argument to request smooth data flows could be helpful to some extent, particularly to research groups that have the possibility to allocate resources to engage with groups of patients on a regular basis, and therefore feel confident when presenting the argument that they have the support of these groups. However the fact that some data providers still show reluctance to share data suggest that for them to agree to share data it is not enough to know that university-based researchers have the support of patients and that what they want to do is legal. As shown by previous research, people working at healthcare institutions consider it important to comply with legal frameworks and take into account the views of patients and members of the public, however making legal uses of data and having the support of patient groups and members of the public is not enough (Neves et al., 2019). Before supporting the reuse of patient data for a specific project practitioners working in the healthcare sector expect to be reassured that data is going to be used in an ethical way and that outcomes of research will not damage patients, specially those who are most vulnerable.

7.2 The patient and public support as a source of legitimacy

Having the input from the public and patients allows researchers to understand what uses of data are acceptable for these two groups of stakeholders, but it also brings a key additional benefit for them. Some participants of this research expressed that they are aware that having the support of patients does not excuse them from following the regulations that establish the ways in which patient data can be used in legal and institutional terms, nor grants them the

authorisation to break data protection laws. Nevertheless, having the support of patients and the public is helpful in the sense that it gives them something that seems to be very important for them, especially following the Care.data scandal. Engaging with patients and groups of members of the public gives the researchers the confidence in the legitimacy of their own actions.

For example, a member of the *Understanding age inequalities and inequities in the cancer pathway* team (Data Journey 5) explained that since their project is reusing patient data without asking for the explicit consent of patients, he and his team felt compelled to take into account the views of patients in order to legitimise their uses of patient data. This participant also stressed that he does not envisage conducting this project without the input of patients and without communicating how they are using patient data:

> We feel we don't have any choice but to have as much input as we can from the patient community for reasons of legitimacy that's also why we have to communicate what it is we're doing. (S4-LS-2)

Similarly, when asked what is the value their team perceive in having a patient group, a participant commented:

> [Without a patient group] we will be a lot more worried because how do you know you are doing the right thing with the data in the first place? The patient voice gives you not legality because you can't break the law, it gives you legitimacy, it gives you the confidence you know, somebody said that you are doing the right thing, that you are doing a good thing. (S7-GN-2)

The voice of patients is considered valuable for participants of this research, not only because this allows them to understand what uses of data are acceptable for these two groups of stakeholders, but also because the support of patients gives confidence in the legitimacy of their research projects.

Whereas the majority of participants from research teams that have regular interactions with groups of patients (e.g. DJs 2, 3 and 5) expressed that having the support and endorsement from patients was important for them, a small number of participants expressed a contrasting view, as they believe that having the agreement or endorsement from patients should not be necessary and that patient data should be used with or without the support of patients:

> I think well, the data was generated by NHS care, so it was generated under the healthcare system so it strikes me to then say you can't use that information to help other people. It's a bit bizarre to me is just a logical step then use that information to try and help other people in the same way you taking taxes of somebody else to give that person treatment what strikes me is very odd that then you say you can't UK healthcare data that is generated from it. To help other people to say that to me is bizarre so it strikes me that should be the presumption that the information will be used and share with appropriate safeguards that's so for me you shouldn't be about asking people are you happy with your healthcare data is used for research that should just be a given that my personal opinion where the emphasis should be is that data appropriate appropriately protect to me that's the important question not the people views. (S4-LS-2)

Perceptions similar to the one above however, were less popular among researchers, most participants across many of the data journeys felt that they needed to hear the voice of patients, even if they are small groups or not representative because this makes them feel like they have approval, that their work is accepted and even endorsed by patients.

7.3 Having the support of groups of patients and the public has helped to make university-based researchers feel confident in the legitimacy of their work, and therefore has motivated them to request access to use patient data.**Patient groups as an opportunity to engage in meaningful conversations and debunk misconceptions around the used of patient data**

Some participants who have experience working with patient or public groups from the *Building Rapid Interventions to Reduce Antibiotic Prescription* (Data Journey 2) and *Understanding age inequalities and inequities in the cancer pathway* (Data Journey 5) teams commented that from conversations with patients and the general public, they have learned that it is common for people to have misconceptions around the uses of patient data. They mentioned for example that in general, people are not familiar with what information is recorded in their electronic health records, and with the type of data researchers have access to when conducting research. One of the most common misconceptions that interviewees recall from their interactions with patient and public groups is the belief that researchers have access to all patients' personal information and details about their care they receive as well as medications. For instance a researcher working on *Building Rapid Interventions to Reduce Antibiotic Prescription* (Data Journey 2) expressed:

> People don't understand what data we got and they think the whole patient record and they think you are getting all the notes, and exactly what you know, what they had for breakfast and in full detail and that is not what we get, we don't know anything about the details of their care, we do know about details of their care but we don't know, no at the level of personal detail they think we got. (S1-YR-1)

According to some participants of this research, misconceptions like the one mentioned above can provoke discomfort and distrust among patients and the public in relation to the idea of using patient data for secondary purposes. A member of the *Understanding age inequalities and inequities in the cancer pathway team* (Data Journey 5)commented:

> There was one lady who didn't have chemotherapy because she didn't want her data to be used yeah and they have proper arguments about what we got and what we haven't and how we should be handling it which is brilliant cause that is exactly what we want but we gotta get let them decide and come to a consent about what is going on, that is exactly what we want. (S4-AS-1)

This suggests that university-based researchers have a good understanding of the levels of knowledge regarding the use of patient data within groups of patients and the public. As previous research has pointed out, people in general have limited knowledge regarding the uses of patient data, and in fact not only about the use of patient data (Understanding Patient Data, 2018), but the uses of all types of personal data (Hartman et al., 2020).

In the view of some participants from some data journeys who have had interactions with groups of patients, the patients who do not support the reuse of patient data for research purposes have this stand because they have not received a clear explanation of what the reuse of patient data entails, and that in their opinion most of the people who understand the implications and benefits of data sharing do not have an objection in relation to the reuse of patient data for secondary purposes:

> I don't think it is really well explained to patients and I think most patients would be quite happy to have most of their data shared for non-commercial purposes. (S4-LS-2)

The view that when people receive a clear explanation about what data sharing is, they are more likely to support and endorse data sharing initiatives was shared among several participants. For example, one participant who have several years of experience engaging with groups of patients commented that for him communication makes a great difference, and that he knows by experience that people at the beginning can feel sceptical about the idea of reusing data, but once the myths are debunked, they change their attitude and express they are happy to share data:

> Once you explain to people exactly what we got and you are clear about it then they are like "well why wouldn't you use that". And again that is why our patient group is important, as soon as they are clear on the data you got and how we are gonna use it they are like "please use it". So I think is all about better communication. (S4-SS-4)

Similarly, another participant who engages on a regular basis with patients reported:

People still think that we are talking about their physical medical records, and they think we are looking at their primary care paper records or something, and is like just even just showing them that actually is a spreadsheet with numbers and you cannot tell anything from that, then that makes a difference. (S2-FA-1)While participants are aware of the levels of understanding and knowledge from people regarding the uses of patient data they do not seem to have a very clear understanding of the reasons why patients and the public sometimes are against or refuse to share patient data. As seen in the quotations above, they seem to believe that the main reason why people refuse to share data is because they do not know exactly how data is going to be used, but once they receive an explanation they should be willing to support the reuse and sharing of data. As previous

research has shown, while some people can change their mind and become more supportive of reusing patient data once they receive information, not everyone has the same reaction, as some people can become more worried or wary of data uses if they know more about how data is used (Connected Health Cities, 2017). A recent study has also demonstrated that people who know more about data tend to become more worried about how their data are used (Kennedy, et al., 2021).

Convincing people about the benefits of data sharing, debunking myths and explaining what the reuse of data is takes time; according to some participants from different data journeys who have closely work with groups of patients, meaningful conversations with participants require time and effort as people cannot be convinced in a short conversation:

> For a society point of view because things that you mentioned that people get their scare and the perception is oh is awful but you have to have the conversation and that conversation will take time is not a 5 minute conversation because people awareness will be no you can have it, the end to be more risk-averse until you tell them what's sharing and what linking. (S4-NS-3)

Participants commented that patients and the public can show more acceptance of secondary uses of patient data when they are provided with examples to which they can relate, and when they receive explanations about the direct positive impact that sharing data can have on the lives of others. For example, a participant commented:

> You are going to these meetings and speak to patients and you see how they speak about how they want their data to be used and then you know "I don't want some of my experiences to happen to my grandchildren if they end up being diagnosed with this and if we could prevent that by finding the stuff in this data then I really want that to happen. (S4-YS-5)

Some participants of this research expressed that they do not understand why people feel comfortable with sharing their personal data in other ways, such as social media, but do not feel comfortable with reuse of patient data for secondary purposes:

> I find peculiar that people have an aversion to sharing their health records or have an issue when they actually if I had access to all of your kind of shopping data and mobile data I would be able to predict way better when you gonna have dementia cause there is a little snapshot where you interact with the system every four years is a lot less useful than a second by second touchpoint with your mobile phone or wherever. So I find ironic that people get upset about medical information yet. (S3-EN-1)

According to some interviewees who have participated in efforts to communicate with groups of patients and members of the public, engaging with patients and the public adequately requires commitment as it is not a simple task. Public and patient engagement "is good fun but is also hard work"; as a participant commented, it requires coordinated work and the engagement of different members of the research project team.

Engaging with patients and the public requires much more than inviting people to attend public events; it is building a relationship based on trust and maintaining open channels of communication for people to reach the research team if a doubt emerges or if they want to know more about any aspect related to the secondary uses of patient data. For example, a participant commented:

> Patients at the heart of everything, which a lot of people say and nobody really does. The idea of patient engagement for some people is inviting them to a meeting, put them in the corner of the room with a cup of tea and some biscuits, and then at the end of this they say thank you, have you got an expenses form? That's it, and I think that is just disgraceful, is disgraceful that that happens. (S4-SS-4)

Apart from organising events with patients, research groups have adopted other ways of communicating with patients the benefits of using patient data for research purposes. For example, one participant explained:

> I lead this project about developing animation. So yeah, definitely part of the reason we are doing this animation is to make sure that we are trying to build up the knowledge of the general public of what data linkage is routine data in the context of research and anonymisation, privacy, you know how is data gonna look like. (S2-FA-1)

At the present time research groups are increasingly engaging with patients and the public in different ways; however as some participants commented, not all research teams are willing or interested in doing this. According to participants on many data journeys, this may be because doing so is really hard work, for example a member of the *Understanding age inequalities and inequities in the cancer pathway* team expressed:

> So yeah I understand people's hesitation to do it, it can be quite daunting. We get people to work very hard to keep the interaction open to do it properly but I think is worth for sure. (S4-YS-5)

The above quotation suggest that university-based researchers are aware that explaining data practices to people it is not an easy task (Centre for Data Ethics and Innovation, 2020).

Previous research suggest that the public and patients support the reuse of patient data for research purposes, however this support is not unconditional. As discussed in Chapter 2 people are only willing to support the reuse of patient data for research purposes if data are used for the public benefit and if if confidentiality can be maintained.

While at present time the voice of patients and the public have not acted as significant barriers for data to flow to universities, if adequate measures are not in place, and if not enough efforts are made to make data practices more transparent people could stop supporting the flow of data towards these organisations as has happened in the past with badly handled initiatives (Hays & Daker-White, 2015). Therefore, the engagement of university-based research groups with patients and the public might be one of the factors that could be helpful to prevent potential

frictions that could emerge if people show opposition towards the circulation of data towards universities.

7.4 Chapter conclusion

This chapter explains the different strategies used by university-based researchers to rebuild the trust that according to their perceptions was lost after Care.data.

University-based researchers have tried to use the voice of patients to convince staff at sites of data production to share patient data with them. This argument however, has not always been useful as for some people at sites of data production this argument is not enough to enable smooth data flows towards university research groups or projects.

Some participants commented that after Care.data the community of university-based researchers engaged in a deeper reflection about how they had integrated the views of patients and the public in their research. They have also designed and implemented strategies to integrate the views of patients and the public in their projects. In these strategies they have make an effort to convince these groups to endorse and support their research by providing explanations about their data practices (how and for what purposes they plan to use patient data).

The work that they are doing to prevent fricitions in data flows seem to be motivated by their desire to overcome frictions generated at sites of data production and preventing the emergence of new frictions that could emerge if patients are not happy about data uses. The engagement with patients and members of the public suggest that researchers are aware that the acceptance of patients and the pulic is key and that as previous research has demonstrated scandals can have a negative impact in the preceptions of the public regarding patient data uses (Healthwatch England, 2018).

While these efforts are helpful, it would be useful for university based researchers to pay more attention to the motivations of patients for supporting the sharing of patient data. As previous research has pointed out, while some people can become more supportive of uses of patient data as they know more, not everyone reacts in the same way, as other peope can become even more worried and therefore less prone to share data (Tully et al., 2018). It would be also useful for university-based researchers to pay greater attention to the fact that the opinions of some groups of patients and members of the public are not necessarily the opinions of everyone; as

research has demonstrated not all groups share the same concerns and expectations (Cancer Research UK & Macmillan Cancer Support, 2016; Kennedy, et al., 2021).

As explained before, university-based researchers seem expect to be granted permission to access to patient data without constraints because they see themselves as an ethical group that will always conduct research in the public interest/to benefit the public, this in interaction with the support of some groups of patients and members of the public for conducting research with patient data has generated different effects. On the one hand the support from these groups has served in first instance to give researchers the confidence in the legitimacy of their own actions. On the other hand, while researchers have tried to lubricate the frictions to access patient data generated by the resistance of healthcare institutions to share patient data by presenting the argument that they have the support of patients and the public, this has not always been useful for that purpose. This is because professionals at healthcare organisations do not only consider important the support of these groups, they also expect to be reassured that data is going to be used in a secure way, following ethical principles and guaranteeing that patients are not going to damaged, specially those who are most vulnerable.

This has helped driving data flows towards their groups as this has been useful to prevent potential opposition from these groups to their data practices.

Chapter 8. Discussion

8.1 Introduction

As stated in the introduction of this work, the aim of this research is to gain understanding of how sociocultural values and norms interact with existing and emergent material conditions to shape "data journeys " (Bates et al., 2016) of patient data in the UK health sector, with a specific focus on reuse of patient data for research purposes. It examines journeys of data reused for research purposes in universities. The research questions that are addressed by this study are presented as follows.

Main Research Question:

- **RQ1** How do sociocultural values interact with existing and emergent material conditions to generate drivers and frictions that work together to produce the material forms of data journeys in the context of reuse of NHS patients' personal data for research purposes?

Sub-Research Questions:

In order to answer the main research question, the following sub-questions are posed:

- **RQ2** What material forms do the 'journeys' of NHS patients' personal data take between different sites of practice, from their initial generation though to reuse for research purposes in different contexts?
- **RQ3** In what ways do sociocultural values interact with existing and emergent material conditions to act as drivers to move data between different sites?
- **RQ4** In what ways do sociocultural values interact with existing and emergent material conditions to act as frictions on efforts to move data between different sites?

This chapter discusses the key findings of this research. It is worth mentioning that in general some of the types of issues that the participants drew attention to reflected key themes identified in techno-managerial and health literatures concerning data flows. For example, participants expressed that technical and data management are major barriers that can slow down or block data movement as they flow between people, organisations, or machines. They also highlighted that datasets produced in the healthcare sector are often incomplete, fragmented and inaccurate,

and expressed a strong interest in engaging in the development of strategies that could help to overcome barriers to data flows. In this section, these findings will be discussed through the lenses of the Critical Data Studies approach.

This chapter is integrated by five sections. Section 8.2 explains how the embracing of the big promises of big data has led to the emergence of a number of factors that act as drivers for patient data generated in the UK healthcare sector to flow to universities. Findings of this work show that the harmony in the views held by university-based researchers and key stakeholders about data and their potential for improving health outcomes and advancing health research have acted as a driver for data to flow from the healthcare sector to the hands of university-based researchers. Section 8.3 discusses the strong views that university-based researchers hold about their identity and culture and how they see other key actors in the healthcare research data landscape: researchers from the pharmaceutical industry and data providers. University-based researchers have constructed and projected a virtuous self-identity before external stakeholders. Some funders, data providers and policymakers have endorsed such virtuous self-identity and this has fostered the flow of patient data to universities. However, this self-identity has not always been helpful to smooth frictions to negotiate access to patient data with staff at sites of data production. Section 8.4 discusses university-based researchers' perceptions concerning the negative impact on their research endeavours such as Care.data and other controversial data-sharing partnerships between the NHS and external stakeholders. This section also discusses the efforts that the community of university-based researchers has undertaken to regain public trust and the impact that this has had in shaping patient data flows. Section 8.5 discusses the infrastructural barriers that have slowed down or blocked the movement of patient data from the healthcare sector to universities for conducting research. Finally Section 8.6, explains how the different factors explored in this section work together to shape patient data movement.

8.2 The promise of data

University-based researchers have expressed in their discourses that they see data as a resource with 'great potential' to improve health outcomes and advance health research; therefore, they hold grand expectations about them. The ideas expressed by university-based researchers seemed to be aligned with views expressed in the discourses and institutional priorities of key stakeholders in the healthcare and healthcare research landscape. The embrace of big data promises by university-based researchers, combined with the alignment of views between this

community and key stakeholders in healthcare and healthcare research, has contributed to facilitating the flow of patient data from the healthcare sector to university-based research groups.

To understand the expectations for and values around data embedded in university-based researchers' discourses, attention was paid to what Fiore-Gartland and Neff (2015) define as 'data valences'. These authors described six ways people within health and data wellness communities talk about data: “self-evidence, actionability, connection, transparency, truthiness, and discovery” (Fiore-Gartland & Neff, 2015, p. 1466). These six valences were present in the interviewees' discourses; however, I will focus on four of them in this section and in the vanguard valence, which is the additional valence I proposed drawing from the findings of this research. Therefore, the valences to be addressed are self-evidence, actionability, discovery, truthiness, and vanguard; this is because they are particularly useful to shed light on the "expectations" that participants hold about data. The transparency valence is addressed in section 8.3.3, where the benefits this community perceives in permitting the flow of patient data generated in the healthcare sector to universities are discussed. Finally, the connection valence is addressed in section 8.4.3 where the interaction of university-based researchers and members of the public is addressed.

8.2.1 Self-evidence valence

Findings of this research reveal that most interviewees across all data journeys tended to talk about patient data produced in the healthcare sector as self-evident (Fiore-Gartland & Neff, 2015), that is to say, as 'premade' resources that require neither work nor interpretation. The majority of interviewees explained that a key part of their work involves analysing data, which requires a lot of time and can slow down the work with data. Nevertheless, when they describe the value they see in data and their potential, this analysis effort is barely mentioned. They tend to highlight that one of the greatest advantages of working with data is that great results can be achieved with little work. As shown in section 5.3 of the Findings chapter, for university-based researchers having access to large volumes of data was equivalent to being better equipped to answer their research questions, as large datasets, in their opinion, provide fine-grained detailed insights. This assumption that more data is equal to better insights has led researchers to see the large volume of data produced in the healthcare sector as a very rich resource that is highly valuable for them. In contrast, they see other types of data, such as clinical trial data, as less useful and less convenient to use. According to them, these data offer "smaller pockets of data,"

thus working with these types of data demands more time and material resources and is not immediately useful in practice.

The notion that data exist in the world and that gathering them leads to knowledge has been welcomed not only within the community of university-based researchers; other key actors in the healthcare landscape have also embraced it. For example, Fiore-Gartland & Neff (2015) identified that this notion was present among communities of technology designers. Stevens and colleagues (2018) also reported the existence of similar assumptions among scientific editorials of the healthcare domain that have often referred to data as something that just needs to "be gathered to be useful or to lead to knowledge." In the words of Stevens et al. (2018), they talk about data as if they were butterflies or other natural resources that can simply be captured.

8.2.2 Truthiness valence

The truthiness valence, “which illustrates how people expect data to comprise a single, direct, objective representation of a measured reality” (Fiore-Gartland & Neff, 2015, p. 1476), was also present in interviewees' discourses. As observed in the findings, most participants talk about patient data as accurate representations of the world. They tended to refer to them as datasets with great coverage that 'give a whole picture, real-life or real-world examples'. They see patient data as 'more objective' and 'truer' than other types of data, such as randomised control data. One of the ideas expressed by several university-based researchers was that clinical trial data have massive gaps because of their limited coverage. The view expressed by university-based researchers that patient data provide accurate representations of the real world suggests that participants have a tendency to overlook the fact that data production is a result of human action. And it is precisely because of the human intervention that data cannot be neutral and objective or 'accurate representations of the world' (Edwards, 2013; Ribes & Jackson, 2013). Assuming that data can depict neutral and objective representations of the world is problematic for two key reasons: firstly, it does not recognise that data is always an oversimplification of reality. For that reason, certain aspects of a phenomenon might be emphasised while others might be obscured (boyd & Crawford, 2012). It is also problematic because it ignores the social and political processes involved in creating data, which cause them to be subjective and infused with biases (boyd & Crawford, 2012).

8.2.3 Discovery valence

Another valence identified in the discourses of interviewees was the discovery valence. This valence depicts how people expect data to be the genesis of discovering phenomena, issues, relationships, or states that otherwise would remain unknown (Fiore-Gartland & Neff, 2015). The accounts of the majority of university-based researchers interviewed were in agreement with the logic behind the discovery valence, which places great importance on finding patterns. A strong tendency to highlight the usefulness of finding patterns in data was identified among participants, as if this was equivalent to knowing or understanding patterns in health issues at population scales. In addition to this, most of them expressed interest in conducting research using patient data because they would not have another way to discover things without access to those pockets of data. Whereas most participants commented that "many things" could not be discovered without having access to patient data, they did not give specific examples of those things that they were expecting to discover. Therefore, it was not clear whether they were hoping to find new diseases, answers to specific research questions, or only patterns. However, most of them expressed a sense of urgency and a strong interest in conducting research using data. They expected that data would be the source for discovering what otherwise would remain hidden or out of reach.

This excitement around the potential things that can be discovered through data is not exclusive to the community of university-based researchers and data practitioners. Similar views are held by actors in other contexts, such as commercial organisations and different sectors of the government (boyd & Crawford, 2012; O'Neil, 2016). For example, within the UK, an increasing emphasis on data use in government has led to a "proliferation of data systems being implemented, leading to significant experimentation with algorithmic processes designed to provide new insights and value extraction based on different kinds of analytics" (Dencik, Redden, et al., 2019, p. 19). Findings of this research suggest that most university-based researchers perceive data-driven research as an innovative opportunity to positively influence health outcomes. They spoke with a sense of urgency about the need to continue conducting and incentivising health data-driven research, pointing out that they should not be left behind if other sectors are doing it. Despite alluding to "other sectors" in their examples, they only refer to big retail and technology companies such as Amazon, Tesco, and Google.

Discovering patterns to predict outcomes

Whereas most participants spoke positively about the application of data science techniques in general and expressed excitement about the prospect of taking advantage of the technology and the power of computers, they tended toward a particular interest and excitement for predictive analytics. This excitement around predictive analytics has been previously identified in the healthcare domain, but it has also been identified across sectors (boyd & Crawford, 2012; D'ignazio & Klein, 2020; Mayer-Schönberger & Cukier, 2013). On the one hand, predictive analytics have offered important benefits, as they have been useful, for example, to support conflict prevention and identify how refugee crisis might unfold (Mayer-Schönberger & Cukier, 2013). However, on the other hand, their application in many cases has led to disgraceful consequences, for example, the bad handling of "predictive policing" (O'Neil, 2016) systems, which were implemented to prevent crimes from happening, has led to discrimination against certain groups based on race and to "guilt by association".

Over-relianceon predictive analytics is not recommended as predictions are merely estimations of future events expressed in likelihoods; they do not offer certainty. In addition to this, they are often biased and frequently based on fragile understanding. An example of how fallible these systems are is the case of Google Flu Trends, which was released in 2009 when H1N1, a new flu virus, was discovered. Google Flu Trends managed to successfully "predict" the spread of H1N1 in 2009 in the United States by analysing Google searches of users (Mayer-Schönberger & Cukier, 2013). After this, it was believed that Google Flu could be a good tool to predict and prevent the spread of future flu outbreaks; however, this was not the case (D'ignazio & Klein, 2020). In 2012-2013 flu season, Google estimated there would be more than twice as many flu cases as there were. It is not clear if the discrepancy between Google predictions and the actual numbers was perhaps due to media panic about swine flu, or to other unknown factors. However, what this case demonstrates is that it is not a good idea to prioritise prediction and utility over causation and context (Lazer et al., 2014).

8.2.4 Actionability valence: Data Save Lives - data leads to knowledge, knowledge leads to change

The actionability valence, defined by Fiore-Gartland and Neff (2015, p. 1474) as "the expectation that data drive or do something within a social setting or that data can be leveraged for action", was identified in the discourses of university-based researchers and data practitioners interviewed for this research. When participants explained why they were interested in using patient data, one of the most popular reasons was that "data save lives," a

phrase that alluded to the actionability valence. This phrase, heavily used by university-based researchers and data practitioners, is the name and slogan of a public engagement campaign originated at the University of Manchester's Health eResearch Centre in 2014. This campaign aimed at highlighting the positive ways that patient data is securely reused to improve health services (Data Save Lives, n.d.). Data Saves Lives has now been adopted by several research networks and stakeholder groups within the UK and other nations such as the United States of America, Australia, and Denmark. Most university-based researchers and data practitioners interviewed for this research are part of research groups that have adopted Data Save Lives.

According to most participants' accounts, data save lives because previously unsolved problems and unanswered questions can be addressed with data. At the core of these statements is the assumption that having knowledge or answers always leads to actions. In the specific case of the research they conduct, university-based researchers seem to assume that patient data leads to knowledge and that knowledge leads to actions that will always positively affect the healthcare sector. According to most participants, thanks to the use of patient data for health research, better health interventions and treatments are designed, and the health of the whole country is improved.

8.2.5 The vanguard valence

An additional valence, distinct to those proposed by Fiore-Gartland and colleagues (2015), was also identified in this research. The vanguard valence illustrates how university-based researchers and some key stakeholders in the healthcare landscape perceive conducting research with data as the most innovative way of exploring health issues. University-based researchers tended to evoke the vanguard valence when they talked about conducting research with large patient datasets using data analytic techniques. They often highlighted that this type of research positions them as part of a group breaking with old ways of doing science. For example, in the words of a researcher who worked in *Building Rapid Interventions to Reduce Antibiotic Resistance* (Data Journey 1), one of their main motivations to join this project was that this seemed an opportunity to do “something more novel and exciting”.

8.2.6 The promises of big data and their contribution to the socio-material constitution of data flows

Based on the above discussion of data valences, it can be suggested that a key factor that has helped to make patient data generated in the UK healthcare sector flow to the hands of

university-based researchers and data practitioners is that this community seems to have embraced the "big promises of data."Participants tended to highlight the positive aspects of data-driven research, paying little attention to negative issues associated with the use of patient data for research. In their discourses, they talk about patient data, as a premade resource that require little work or interpretation and that comprise a single, direct, and objective representation of reality. In their accounts, they also refer to patient data, as an avenue for discovery, as a tool for understanding patterns of diseases. Interviewees also expressed a strong interest and excitement for conducting health research using patient data.

Some of the interviews have previously conducted other types of research, such as clinical trial research or large qualitative studies. However, they felt motivated to join research projects that involve the reuse of patient data, driven by the idea that this was an opportunity to innovate and become part of a community of visionary scientists conducting cutting-edge research. For this same reason, some of them have sought opportunities to develop new skills or improve their skills in working with large patient datasets.

The embrace of big data promises has helped generate a demand for patient data to flow from the healthcare sector to university-based researchers. However, this fact does not operate on its own. Instead,this factor interacts and works in combination with another one: key stakeholders in the health and health research sector in the UK have also embraced these promises. This is discussed in the following section.

8.2.7 University-based researchers and key stakeholders: a shared view about the potential of data

University-based researchers' opinions concerning the potential of patient data seemed to be in harmony with ideas underpinning the agendas of key organisations that influence how data produced in the healthcare sector are used. As explained in the previous section, self-evidence, truthiness, discovery, actionability, and vanguard are the four key valences identified in university-based researchers' discourses. These valences also have a strong presence in key institutions' discourses in the health and health research landscape. Bodies such as the Department of Health (DoH), the Medical Research Council (MRC), NHS Digital, and the National Institute for Health Research (NIHR) have expressed in varied ways ideas that reveal their embracement of the big promises of data. This strong presence is reflected in the public discourses of their senior staff as well as in organisational visions and strategic statements. Whereas university-based researchers heavily alluded to the truthiness valence, in the

discourses of the bodies mentioned above, the truthiness valence seems to have a weaker presence. On the other hand, a strong presence of the following valences was identified in both groups of stakeholders: self-evidence, actionability, discovery, and vanguard.

The actionability valence is manifested in that all these bodies have promoted and supported the idea that patient data would lead to achieving the ambitious goal of improving the population's health and producing better health outcomes. It is also worth mentioning that all these stakeholders have also been great supporters of the Data Saves Lives initiative, which, as mentioned before, aims to highlight the benefits of using patient data for research.

These bodies have also expressed excitement concerning a key idea at the core of the discovery valence, which is that large patient datasets have an extraordinary power to catalyse discovery. They have also tended to evoke the vanguard valence when explaining their goals and initiatives, and often expressed their interest in creating innovative solutions to tackle health issues. Finally, it is clear that in their discourses, the self-evidence valence is also present. The presence of the self-evidence valence is shown in that they have repeatedly affirmed that diseases can be better understood and new ways to cure and prevent them can be discovered by making patient data available.

The Department of Health, the Medical Research Council (MRC), NHS Digital, and the National Institute for Health Research (NIHR), have all promoted and embraced the big promises of big data. The embrace of the big promises of big data is not only evident in their discourses but also in the actions that they have taken. The Department of Health and Social Care, for example, has expressed its desire for "unlocking patient data" on the grounds that "data save lives" (Hancock, 2019). According to a public discourse pronounced by Matt Hancock in September 2020, one of the key actions this organisation has taken is embarking on reforms to lead to a radical change of culture within the healthcare sector (Hancock, 2020). This change involves, according to him, “embracing patient data as an asset, rather than something that needs to be protected and hidden away”. This view that patient data is an asset that should be exploited has been quite controversial, especially because it has been perceived that this view has benefitted certain stakeholders in the private sector, such as big pharmaceutical or technology companies (Powles & Hodson, 2017).

The Medical Research Council has expressed an interest in harnessing information within clinical datasets to gain new scientific knowledge and has taken a number of actions that reflect that this organisation has embraced the promises of big data. For example, creating policies to

promote data discovery, access, and sharing; supporting cutting-edge research using large datasets; building capacity in skills that are crucial for analysing large datasets; providing financial support for the development of infrastructure, tools, and technology to allow data gathering, storage, and analysis (Medical Research Council, n.d.-b); and working in partnership with other organisations to invest in a various initiatives in data science research. One of the most prominent of its partnership initiatives is the establishment of Health Data Research UK (HDR UK). HDR UK, an initiative aimed at "uniting the UK's health data to allow discoveries that improve the lives of people" (Health Data Research UK [HDR UK], n.d., para. 1), was established through a partnership that includes several organisations: the Medical Research Council, the "British Heart Foundation, the Chief Scientist Office (Scotland), the Engineering and Physical Sciences Research Council (EPSRC), the Economic and Social Research Council (ESRC), Health and Care Research Wales, National Institute of Health Research (England), and Wellcome" (Health Data Research UK [HDR UK], n.d., para. 1).

HDR UK has highlighted that by providing access to rich health data, they aim to understand diseases better and discover new ways to prevent, treat, and cure them (Health Data Research UK [HDR UK], n.d.). One of its key actions is working towards building a cohort to lead the Health Data Science Revolution. Professionals from the following fields are expected to be part of this cohort: biology, research, clinical medicine, computer science, data analytics, and statistics. In addition to this, HDR has set as one of its key goals for the next five years to create more than 10,000 health data scientists to whom they will offer support to become "leaders within health data research" and "to influence the science of tomorrow" (Health Data Research UK [HDRUK], 2019). Finally, they have offered co-funding (alongside with other UK research councils) to individual fellowships whose research involves the reuse of patient data.

Additionally, the main funding body for health research in the UK, the National Institute for Health Research, has communicated its intention of becoming a "truly data-driven organisation," highlighting that it has "just started on a journey to realise its dream of the power of data" (Hirst, 2017, para. 23).

Finally, the NHS Digital, which is the digital partner to the NHS, has communicated that this organisation's role is "to harness the power of information technology to make health and care better"(Ray, 2018, p. 7). This organisation has stressed the need to harness the power of data. Actions taken by it include encouraging the reuse of patient data for research and investing in infrastructural developments to make better use of data, including the building of new platforms, making more data available for researchers, and improving the quality of existing datasets.

The following table condenses the goals and ambitions of the bodies mentioned above. It gives some examples of these bodies' actions that show how they have embraced the big promises of big data.

Table 8. Key organisations in the healthcare landscape: goals and ambitions

Name of organisation	Goals and ambitions	Examples of key actions
Public Health England	"We need to change the culture to see data as an asset that can build the future of the NHS rather than as something that needs to be protected and hidden away…the data needs to be used to save lives"(Hancock, 2020, 9:10)	Proposing reforms with the objective of facilitating the reuse of data produced by the NHS by external organisations (Hancock, 2020).
Medical Research Council	"The MRC's vision is to harness information contained within clinical, population, cellular and molecular datasets to gain new scientific insights into health and wellbeing" (Medical Research Council, n.d.-a, para. 2)	Creating policies to encourage the use of patient data in research
		Providing funding for infrastructure and technologies to facilitate the collection, sharing, and analysis of patient data (Medical Research Council, n.d.-a).
		Investing more than £90m, in partnership with other bodies to support a number of data science research projects (Medical Research Council, n.d.-a).
		Working in partnership with bodies such as the British Heart Foundation, the Economic and Social Research Council (ESRC) to provide an initial funding of £54 to establish Health Data Research UK, created to bring together the UK's health data to enable research to benefit the population (Health Data Research UK [HDRUK], n.d.)
NIHR	"We have only just started out on our journey to realise our dream of the power of data. We are actively working to improve our capabilities in this evolving and exciting area... we will not losing sight of the end goal: to use the power of data to save lives and improve the health and wealth of the nation." (Hirst, 2017, para. 23)	Providing access to health data to researchers with the objective of understanding health issues and discovering innovative ways to prevent and address them (Hirst, 2017).
HDR UK	"Our vision is that every health and care interaction and research endeavour will be enhanced by access to large scale data and advanced analytics."(Health Data Research UK [HDR UK], n.d., para. 1)	Creating in the next five years more than 10,000 health data scientists (Health Data Research UK [HDRUK], 2019).
		Offering funding to researchers interested in conducting innovative research using patient data (Health Data Research UK [HDR UK], n.d.)
		Working towards developing the capacity and strategies to accelerate the development of health data science (Health Data Research UK [HDR UK], n.d.).
NHS Digital	"NHS Digital, indeed the whole health system role is to harness the power of information technology to make health and care better…we need to harness the power of data" (Ray, 2018, p. 7)	Generating "richer and broader datasets" (NHS Digital, n.d.-b, para. 12).
		Undertaking efforts to improve infrastructure to make better use of data (Ray, 2018).
		Building platforms to handle large volumes of data faster (NHS Digital, n.d.-b).

Embracing the big promises of big data

As discussed in sections 8.2.1 to 8.2.5, the discourses of university-based researchers and data practitioners reflect the presence of five key data valences: that data are self-evident, actionable, true, a source of discovery, and that conducting research with patient data is innovative. Participants expressed positive ideas and grand expectations concerning the use of patient data for research. The majority of them pointed out that they started and were interested in continuing doing research using patient data because this is a possibility for innovating, advancing health research, and bringing positive results for the healthcare sector. Thus, it can be argued that these ideas connected to the big promises of big data have motivated university-based researchers to conduct research using patient data and this has generated a demand for data.

As explained earlier in this chapter above the five valences mentioned above are also present in the discourses of some key organisations in the health and health research landscape, namely the Department of Health, the Medical Research Council, the National Institute for Health Research, and NHS Digital. Like university-based researchers, all these organisations have stressed the benefits and grand expectations about conducting research using patient data produced in the healthcare sector. Their discourses are infused with an emotive language encouraging the use of patient data for research. They have tended to highlight the great opportunity that is to conduct research reusing patient data. We can see, for example, in their discourses phrases such as "realising the power of data" (NHS Digital, n.d.-a), "realising the big dream of the power of data" (Hancock, 2019), or "revolutionising health research" (Hirst, 2017). Actionability, discovery, self-evidence, and vanguard valences have a strong presence in institutional discourses. However, the truthiness valence has not been alluded to in these discourses as much as university-based researchers' ones. The big promises of big data have occupied a central role in institutional discourses, but also have underpinned the agendas of these organisations and have led them to take several actions in an attempt to foster the flow of patient data from the healthcare sector to universities, such as the provision of funding, investment infrastructural developments, and design of policies in an effort to reduce some of the socio-material barriers to data flows.

The desire and the interest that university-based researchers show in exploiting data would not be enough to make data flow. However, the discourses and actions undertaken by the key bodies in the health and health research landscape, combined with the ideas that the university-based researchers and practitioners community have embraced. As discussed in Chapter 5, the

interaction of these different factors have helped create a socio-material context that is favourable to making data flow from the healthcare sector to the hands of university-based researchers. The increasing demand of data seemed to be not only validated, but also backed up by institutions with financial and material resources, visibility and decision making power. In the first instance, these organisations' discourses have helped reinforce the ideas of researchers about the benefits and advantages of conducting research using patient data. Furthermore, their actions have generated a response from the university-based researcher community. In recent years increasing funding and support for health data research has been a priority. In other words, the efforts of key organisations might motivate university-based researchers who have not yet conducted research reusing patient data to start doing this. They also might have encouraged those who were already working in projects of this nature to continue in this path for three key reasons: this type of research is highly valued, there are resources available for training if they want to improve their skills, and there is funding for this type of project.

University-based researchers have embraced the big promises of big data. Their interest in doing research reusing patient data, however, also seems to have a pragmatic basis. Incentivising and providing resources to help conduct data-driven research is certainly becoming a priority for key players in the health data landscape. Therefore, there is an open call to embrace big data promises,. It would seem that researchers perceive that conducting research reusing patient data means to seize the opportunity to join the health data revolution and receive support in different ways (through funding, training, support, endorsement). On the other hand, they see that ignoring the call to conduct research using patient data would be equal to being left behind.

For researchers in senior positions, joining the health data revolution opens the possibility to lead on projects of this nature and helps to place their research groups in a favourable or privileged position where they have better chances to receive funding. All because they are doing something that is considered innovative, accepted, and lauded by these different bodies. For junior researchers, embracing the promises of big data and continuing research with patient data represents an opportunity to boost their budding careers. Even if they were interested in doing other types of research, they are told that this is innovative, that is the future, and that this is what they are expected to do. They are motivated by these discourses; thus, the most convenient thing to do is seize the opportunity. If the resources and opportunities are available, the option is to embrace these promises, work towards that goal, and join the innovating group.

An additional benefit that conducting research with patient data would bring to both junior and senior academics is helping them produce journal papers, which is highly valuable in the academic research sphere as it is a mechanism to evaluate researchers. All these factors contribute to sustaining the growing demand for data.

The dangers of embracing the promise of big data

Doing research using patient data can lead to positive results, as acknowledged in participants' accounts. Thanks to this type of research, they have been able to conduct original and important research that would not have otherwise been possible and contribute to developing a better understanding of relevant aspects about a number of critical health issues. However, embracing the promises of data without being critical is not ideal

University-based researchers and data practitioners often connected doing research with large patient datasets with ideas such as innovation, talking of this type of research as a possibility to radically transform healthcare. Adopting this position could be dangerous as one can become so fixated on the data, and so mesmerised with the power and promises it offers, that appreciating its limitations can become challenging (Mayer-Schönberger & Cukier, 2013). Another risk is that embracing the promises of big data can lead to the practice of 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). University-based researchers focused mainly on the grand promises of big data, however, some of them pointed out that if badly handled, privacy issues could emerge while conducting research with patient data. Participants did not talk about potential negative societal effects of the applications of big data, but this might be because they were not asked directly about this in the context of this research.,.

The community of university-based researchers and data practitioners would benefit from considering the "cooking process" that produces data; that is to say, to interrogate the context, limitations, and validity of the data under use (D'ignazio & Klein, 2020). It is also crucial to keep in mind that big data can point us towards understanding, but it can still generate confusion or misinterpretation, therefore we must avoid letting “its seductive glimmer blind us to its inherent imperfections" (Mayer-Schönberger & Cukier, 2013, p. 2).

8.3 Us good, them bad: strong perceptions about data cultures

Beyond perceptions about data that helped drive demand for new data flows and reduce barriers to existing flows, university-based researchers also had strong perceptions about themselves and other key actors: 1) research groups based at pharmaceutical companies reusing patient data and, 2) information governance staff at sites of data production. On the one hand, university-based researchers imagine and present themselves as a community of ethical and virtuous people who always conduct research to produce results that benefit the UK population, such as detecting, preventing, and curing serious illnesses. On the other hand, this community seems to have a negative view concerning other key stakeholders. University-based researchers see pharmaceutical companies that are also reusers of patient data as less virtuous groups that are highly motivated by an interest in gaining profit or gain commercial benefits. Concerning information governance staff in the healthcare sector, some university-based researchers, particularly those who have engaged with information governance staff in the healthcare in the healthcare sector in negotiations to access patient data and have had some of their requests denied, have sometimes perceive these actors as very anxious people driven by fears and anxiety, and who are not interested in sharing patient data and have the potential to hinder researchers' work. In this section, I argue that researchers' virtuous self-identity has helped to some extent to make data flow to universities, not only because university-based researchers present themselves as virtuous but also because some key data providers endorse this identity. However, access to patient data has also been granted to pharmaceutical and tech companies, despite these actors not being considered virtuous as university-based researchers are.

8.3.1 Self-identification as a group led by strong ethical values

As this study shows, interviewees consider that an ethical culture prevails within the community of UK university-based researchers and data practitioners. Whereas they did not use the adjective "ethical" when describing themselves, most of them refer to ethical values when talking about their work. University-based researchers expressed that fairness, transparency, responsibility, and trustworthiness are the core values that underpin how they behave and how they conduct research. According to them, their core values are reflected in their day-to-day research endeavours. These core values are evident, from their perspective, in that their main driver for conducting research is always to improve health outcomes for patients and the UK general population. They also shared the belief that their values are also reflected through the fact that they work to the highest standards that guarantee their excellence in

research. In addition to this, most of them stressed that they handle data excellently as they follow all the necessary information governance frameworks, ethical principles, and regulations to manage patient data that ensure data is safe and secure.

The way university-based researchers talk about their work suggests that this community's actions might be driven to some extent by moral concerns; that is to say, a form of virtue ethics, a virtue of trying to pursue something good (MacIntyre, 2007; Powell, 2019). Virtue ethics is concerned with the process; the focus is not on the outcome of actions, but in trying to pursue something that is good or trying to become good (Powell, 2019). The evocation of virtuousness has also been identified across tech cultures, where technologists hope to do the right thing (Powell, 2017). It has been observed, for example, that within these cultures, virtuousness is often seen in projects that evoke "hacker ethics" whereby technologists have shown a tendency to justify their actions saying that they are driven by "doing the right thing" or “not being evil”. Virtuousness, in this context, has at times motivated opposition to regulatory action. In the case of university-based researchers and data practitioners, the intention of wanting to be doing the right thing, doing good, or not harming was often evoked, which reveals that a form of virtue ethics is a key part of the culture within this group.

8.3.2 The culture of university-based researchers versus the perceived culture of researchers from pharmaceutical industry

Participants' accounts revealed not only that virtue ethics is a key part of the culture of university-based researchers. They also showed that they believe that these strong desires for doing the right thing and doing good do not have a strong presence within the cultures of other stakeholders such as pharmaceutical and technology companies which are other actors that can obtain access to patient data.

The majority of university-based researchers interviewed expressed that their main motivation to conduct research is to benefit patients and the general population. Conversely, research groups affiliated to the pharmaceutical industry or technology companies, in their opinion, will always have other priorities and motivations, such as generating profit or gaining competitive advantage; therefore improving healthcare is not their primary concern. They pointed out that university-based researchers and data practitioners are open and transparent in communicating their research objectives and how they work with data. However, pharmaceutical and technology companies have been characterised by a strong tendency to hide or make it difficult

to access this type of information. Consequently, their real motivations and modus operandi are "grey" and, at times, difficult to decipher.

Participants expressed concern and disapproval concerning some pharmaceutical and technology companies' practices and questioned the ethicality of such practices. During the interviews, they refer to some controversies that have surrounded these companies in recent years. They also highlighted that these companies had been frequently associated with unethical practices such as misuse of patient data and opacity in revealing their research objectives or uses of data.

As told by the participants, university researchers seem to embrace their virtuous self-identity. They seem completely convinced that their research community is a very ethical one. By saying that they can ensure that they work with excellence at all times, the community of university-based researchers and data practitioners appeared unwilling to recognise the flaws in their university research system. They repeatedly asserted that data misuses could not occur within their community because they always work to the highest standards. However, they found it difficult to take such a definitive stance or assume a clear position about pharmaceutical and technology companies, despite considering them less virtuous actors. According to them, research groups based at pharmaceutical or technology companies are prone to act in spurious ways, play nasty, and alter the research results in their favour. Still, on the other hand, they noted the undeniable fact that they can also provide solutions to health problems even if they profit. Also, they believed that their contribution may be on a larger scale than university-based researchers due to the resources they have at their disposal. For this reason, university-based researchers found it difficult to decide whether it is fair or not to give access to patient data to actors in the private sector.

8.3.3 The expectation of frictionless data flows

As discussed in chapter 5 (Section 5.2.2), university-based researchers seemed to expect that they should obtain access to patient data with little or no constraints because of their virtuous self-identity. This expectation became evident in most interviews when participants evoked what Fiore-Gartland and Neff (2015) define as the transparency valence. According to the proposers of this term, this valence is evoked when people talk about the benefits of “making data accessible, open, shareable, or comparable across cases and contexts” (Fiore-Gartland & Neff, 2015, p. 1475). Most university-based researchers explained that they expect to have access to patient data with few or no constraints. According to them, this should happen for

two key reasons: 1) because their ultimate goal is to benefit the population; and 2) because they conduct excellent research. In their opinion, their good intentions and desire for contributing to improving health outcomes should be an acceptable and solid justification for them to get access to patient data.

Participants emphasised that their data requests should be evaluated and treated in different ways because of the differences between them and researchers from pharmaceutical or technology companies. For them, it seems unfair to be evaluated under the same criteria as pharmaceutical or technology companies as their approaches are very different and they have behaved differently. From the perspective of university-based researchers, since they are part of a community with strong ethical values, they should obtain access to patient data with few or no constraints. They also seemed to believe that the university internal review processes used to assess their research projects are rigorous enough to demonstrate that their intended uses of data and methods are ethical and lawful. Their self-evaluation and having the approval and endorsement of their peers on university ethics committees make them feel convinced that there are no reasons for their projects not to be carried out. The self-identification of university-based researchers as virtuous people has motivated them to express discontent with the idea of having their data requests surveyed by stakeholders outside their guild, such as evaluation committees set up by data providers. This is a point about I will talk in more detail in the next subsection.

This self-conception of university researchers and data practitioners as a virtuous community has led them to think that they should have a seamless flow of patient data produced in the healthcare sector. In contrast, they believed pharmaceutical and technology companies should be subject to more strict evaluation processes. In this way, they would be prevented from making incorrect uses of patient data because they are not as virtuous as them and are more prone to having 'spurious goals' and are willing to cheat, lie, or not be transparent.

Nonetheless, whereas participants reject the idea of being evaluated in the same way as actors from pharmaceutical and technology companies, they are not against working in partnership with these actors. They seem to be willing to collaborate with them because they feel convinced that working in collaboration does not entail the same risks for patient data as granting permission to these companies to work on their own. They seem to feel that if they are a presence in the room then they are capable of preventing private sector collaborators use data in an unlawful or unethical way, able to overlook collaborators conduct as if they were a purifying entity. They also recognise that pharmaceutical and technology companies can contribute to achieving important goals in health research. Although university-based

researchers showed a willingness to collaborate with pharmaceutical and technology companies if the opportunity emerges, they do not have a clear answer about whether it is fair or not for these actors to have access to patient data. It would seem that this willingness to collaborate is a form of compromise (accepting a bit of them but not giving them data blindly or with hands wide open).

8.3.4 A virtuous identity can open data flows

University-based researchers have embraced a self-identity as virtuous and ethical people and have presented themselves in this way to funders, data providers, and the public. This image has impacted the flow of patient data, as it has helped to make data flow to groups of university-based researchers. In the first place, funding bodies and data providers seem to agree with this image of university-based researchers. Funders and other key bodies have expressed that they value and recognise the university-based researchers and data practitioners community's virtue. Therefore, they have shown that they trust them and are willing and interested in collaborating with them (Vezyridis & Timmons, 2017). This has been demonstrated as both data providers and key funding bodies repeatedly have shown their endorsement of this community. Their support has been shown through the public praise of the culture of ethical and excellent research of university-based research groups and the open call to these groups to conduct research using patient data (Blaveri, 2017). Alluding to the good reputation of ethical and excellent research, key funding bodies have invited university-based researchers to put forward projects to conduct high-quality research with patient data from the UK healthcare sector. It is under the same argument that they have been invited to take part in NHS data sharing initiatives (Vezyridis & Timmons, 2017).

In contrast, key stakeholders have not endorsed pharmaceutical and technology companies' virtuous identity in the same way as they have done with the university-based research community. However, this has not restrained them from granting these companies access to patient data or collaborating with them (Ramanauskas, 2019). In fact, in recent years, a number of partnerships with the private sector have been agreed (Collington, 2019).

As discussed in the first section of this chapter, key stakeholders such as funders and data providers in the healthcare landscape have expressed their interest in exploiting patient data. One of the key reasons underpinning this interest is that they see patient data as assets that can potentially bring financial benefits to the healthcare system. Therefore, working in partnership with private companies has been seen by key stakeholders as a good opportunity

because of the advantages that this offers. From their perspective (Hancock, 2020), these companies offer a great capacity to deliver results quickly, are equipped with material resources such as strong infrastructure and cutting edge technology, and have the ability to manage projects at a large scale. These companies are also perceived to have the ability to offer financial gain. This suggests that whereas key stakeholders have endorsed the virtuous self-identity of university-based researchers, when other interests such as the opportunity of obtaining a financial benefit emerge, working in partnership with 'virtuous' research groups becomes less important and is not a priority.

The virtuous self-identity projected by university-based researchers interacts with another factor which is the support and endorsement offered to them by funding bodies and data providers because they are perceived as an ethical group that conducts excellent research. Therefore it could be argued that the interaction of these two factors has helped to facilitate the flow of patient data from the healthcare sector to university-based research groups. However, patient data flows have also been open to pharmaceutical and technology companies, which despite not being associated with the same virtuous identity, are perceived to offer potential benefits to the public sector.

As explained before, for university-based researchers, patient data-sharing initiatives between the NHS and pharmaceutical or technology companies can have negative consequences. From their point of view, these companies are always driven by a commercial interest, therefore handling patient data adequately and safely is not one of their priorities. In addition to this, they also have the perception that these companies are prone to manipulate results to favour themselves.

Whilst not mentioned by the interviewees, other issues associated with granting access to patient datasets to private actors are worthy of being highlighted. For example, it has been highlighted that private medtech sector actors in the past have made big claims about the results that can be achieved, or in other words, overpromising and underdelivering, not taking into account that having powerful technology is not enough (Strickland, 2019). Partnerships between the NHS and private actors can also be a matter of concern when these initiatives involve not only granting access to patient datasets, but also investing in them to develop technologies instead of public sector research institutions. When this happens, institutional dependencies on these companies to deliver services in the future are created (Collington, 2019). According to Collington (2019), public-private partnerships that involve big data transfer will diminish the power of the public sector in the future. It also could risk more

instances in which public actors are not able to evaluate the claims made by private companies about what results they can achieve or what their products can do.

8.3.5 Staff of healthcare sites of data production perceived as risk-averse

According to the discourses of university-based researchers and data practitioners, another factor that has negatively affected patient data flow, or that acts as a friction for data to move, is negotiations with information governance staff at different sites of data production. They believe that negotiating with information governance staff is extremely difficult, as these actors have shown reluctance to share patient data. We can observe that the community of university-based researchers views the information governance staff as a community that creates obstacles for their "virtuous" goals and objectives.

The findings of this research show the existence of tensions in culture and perceptions between university-based researchers and data practitioners and some information governance staff in healthcare settings. The majority of interviewees have had data requests denied at some point in the course of the last five years (from 2015 to 2020). While data providers had informed them of the reasons why their requests had been denied, a pattern observed in the comments made by interviewees was that some of them in certain cases have attributed the data request refusals to emotional responses from data providers. Some have tended to label healthcare sector staff as "extremely scared", "people who react with nervousness when discussing data sharing", "very anxious individuals" or "not interested in sharing data beyond the healthcare sector".

As explained in a previous section of this chapter, university-based researchers tend to see themselves as a virtuous community, and expect that this self-identity should smooth the frictions to access patient data. The virtuous self-identity has helped them to regain the public trust in university-led research projects which was lost as a result of the Care.data scandal. However, it would seem that their virtuous self-identity has not had the same effect in their negotiations with information governance staff at some sites of data production and this generates negative emotions in them; they expressed feeling frustrated and disappointed by the fact that it is really difficult to obtain access to patient data produced in the healthcare sector.

The fact that they suggest that it should be enough for information governance staff to know that their intention is to conduct research that benefit the population and that what they are proposing to do is legal suggests that they would expect their "virtuous" self-identity would grant them the privilege of free and unquestioned access to data.

Interviewees also pointed out that people in the healthcare sector at times forget the importance of 'their duty to share' as asserted in the 2016 Caldicott Review (Vezyridis & Timmons, 2017). While it is true that the Caldicott report stressed the importance of complying with the "duty to share", it does not refer to the duty to share data for secondary purposes; what it states is that "the duty to share information for the direct care of patients can be as important as the duty to protect patient confidentiality" (Department of Health, 2013b, p. 5). Other studies have pointed out the existence of a risk-averse culture in the healthcare sector; however, these studies refer specifically to a risk-aversion for sharing patient data across different services within the health sector to provide the right care to patients and not to the sharing of data for secondary purposes. The fact that interviewees refer to the "risk-averse" culture within the healthcare sector and expressed that by denying access to patient data for secondary purposes information governance staff are not fulfilling their "duty to share", suggests that perhaps interviewees have misinterpreted what the Caldicott Review and other research really talk about.

Since information governance staff were not interviewed as part of this research, it is not possible to contrast the views of this community with those of the university-based researchers and data practitioners. However it can be suggested that it is possible that university-based researchers and data practitioners, immersed in their discourse of "virtuous" community, have difficulty in viewing weaknesses and potential negative implications of their data practices.

Whereas it is true that the decisions of information governance staff act as frictions for patient data to flow to university-based research groups, these frictions are worth being preserved; and, and it would be relevant to conduct an exploration of them in future research. Despite the expectations of university-based researchers and data practitioners of having unlimited access to data due to their "virtuous" self-identity, it would not be beneficial to open up data flows based on intentions and expectations of a particular group. Additionally, it is important to consider the fact that just because something is legal, does not mean that it is ethical and therefore, does not mean that it should necessarily be permitted.

8.4 Strengthening a fragile public trust

While after Care.data a significant number of patients opted out of their data being used for purposes other than their direct medical care, findings of the literature review suggest citizens are often willing to support the reuse of patient data for research purposes if the research conducted is for the public benefit and if adequate measures are in place to ensure the safe and secure handling of data (Skovgaard et al., 2019; Tully et al., 2018). Nonetheless, participants

of this study have tended to perceive that the levels of trust from patients and the public regarding the reuse of patient data have diminished due to controversial data sharing initiatives between the NHS and external stakeholders. This, from the perspective of interviewees, had generated frictions in the circulation of data towards their research groups. In an attempt to lubricate the frictions generated by a fragile public trust, university-based researchers have adopted strategies to generate support and endorsement from the public. These strategies have often appealed to researchers' virtuous self-identity using as a central argument that "they want to do good" and "do the right thing".

To some extent, some groups of patients and the public have played the role of allies for university-based researchers, as their support has granted researchers the social licence to use patient data. University-based researchers have also used the support of patients and the public as an argument to demonstrate the legitimacy of their work when requesting data providers to allow the circulation of patient data towards their groups or projects. This section discusses the key strategies that universities have adopted to strengthen public trust and how this has helped to smooth the flow of patient data towards their groups or projects.

8.4.1 An interest in restoring trust

As explained in the introduction of this thesis, in 2014, the NHS proposed the release of Care.data, a controversial initiative that involved the secondary use of personal data (Hays & Daker-White, 2015). This initiative generated a lot of confusion because of the lack of information given to the population about what Care.data would involve, what data would be available to whom, and in what form (Kirby, 2014). Some citizens raised concerns regarding a lack of transparency and insufficient respect for confidentiality and privacy (Sterckx et al., 2016). After the controversy caused by the Care.data initiative, its launch was cancelled and frictions around access to patient data were intensified.

It cannot be denied that the lack of approval and opposition from the public were key elements that led to the cancellation of Care.data (Carter et al., 2015; Hays & Daker-White, 2015; Vezyridis & Timmons, 2017). As discussed in the findings chapter, Care.data triggered civic responses from the public aimed at generating frictions to block data flows towards sites of data reuse. For example, when the initiative's controversies became public, around 1.5 million people opted out of their data being used outside of direct care (Vezyridis & Timmons, 2017). Even when Care.data was stopped, those people remained on the list of opt-outs. According to the accounts of university-based researchers interviewed, although they were not involved in

Care.data, this initiative's results reminded them that transparency and taking into account patients' and the public's views are key. Concerns from some members of the public regarding the potential negative consequences that could emerge from sharing data outside the healthcare sector have increased due to controversial sharing initiatives. However, as pointed out in the literature review support for the reuse of patient data for research purposes if this is for the public benefit has remained high (Healthwatch England, 2018).

8.4.2 Appealing to a virtuous self-identity to regain trust

University-based researchers felt that after Care.data they have had to lubricate the friction generated by a fragile public trust. They have attempted to do this by gradually building back public trust around the use of patient data for purposes beyond direct care. To regain the trust of the public and patients, university-based researchers have taken a number of actions. Whereas each research group has adopted different and varied strategies, one key element at the core of these actions is that they have tended to highlight university-based researchers' virtuous self-identity. A central aspect of their strategies consists of engaging in conversations with patients and citizens to inform them how they are using patient data, asking them to provide feedback and share their concerns. For example, one team participating in the study has set up Citizens' Juries with the aim of understanding whether their intended uses of health data are acceptable or not for the public. Other teams have integrated patients or members of the public into their steering committees.

These strategies have strongly appealed to researchers' virtuous self-identity with the intention to combat opposition and to generate endorsement from patients and the public for their projects, with their key argument being that they want to do good and they want to do the right thing. They have expressed that it has been key for them to reassure patients and the public that they work with data responsibly and that their main goal is to improve people's lives and health outcomes for the whole UK population (doing good).

8.4.3 Engagement with the public and patients has helped to make data flow

The engagement of university-based research groups with patients and the public has lubricated the friction generated by a fragile public trust. To some extent, some groups of patients and members of the public have played the role of allies for university-based researchers. This is because establishing a relationship with these groups has helped university-based researchers gain their endorsement, which has been helpful to legitimise their research groups. Gaining

public trust works as a lubricant of friction because this grants researchers a social licence to conduct research using patient data.

When talking about their public engagement initiatives, university-based researchers tended to evoke the connection valence. This valence, according to Fiore-Gartland and Neff shows the expectation that data can be an excuse for engaging in meaningful conversations or an opportunity for making a connection with people (Fiore-Gartland & Neff, 2015). Participants tended to express that public engagement initiatives gives them the possibility to share real life examples of how research with data has benefited other people. According to them, sharing stories to which patients and the public can feel identified generate positive reactions and support. They also pointed out that these initiatives allow them to listen to the fears and concerns of people and debunking myths concerning the reuse of patient data.

Some university-based researchers, mainly those working in well-established research groups have used the support from patients and the public as an argument to prove the legitimacy of their research before data providers and to convince them to share data when they show reluctance. As an interviewee who participated in the project *Understanding age inequalities and inequities in the cancer pathway* (data journey 5) expressed, “the voice of patients” is the “best weapon” that a research team can use when data providers show resistance to share data. An argument that according to university-based researchers' perceptions sometimes can be very powerful to convince data providers to approve their data requests is that having patients' approval is what is essential.

When reflecting on the effect that their strategies have had in rebuilding public trust, some of the university-based researchers interviewed appeared proud of the positive results they have achieved so far. They felt particularly proud of two achievements; first, of having managed to establish what they call a trust-based relationship with their public or patient groups where patients can express their concerns and expectations. Second, of complying with a moral duty of informing the public how they are using patient data. While it is a good idea to listen to patient expectations and concerns, it is not clear why university-based researchers consider this one of their main achievements, as no consent is required for using patient data in their projects. This means that independently of the concerns raised by individuals, some of the existent data flows are unlikely to be blocked. It would seem that university-based researchers think that through their public engagement groups, they are informing "the public"; however, saying this reflects that they might not be taking into consideration demographic variables such as age group, gender, ethnic group, disability, sexuality, and education differences that exist between

different groups of the population. Informing one sector of the population is not equal to informing all sectors of the public (Skovgaard et al., 2019).

University-based researchers and data practitioners have defended their virtuous self-identity in an attempt to smooth patient data frictions, and at the same time, have accused the other "less virtuous actors" such as big technology companies of undermining their efforts. They believe that the controversies around multiple NHS data-sharing initiatives that, according to them, have been badly handled have had a damaging impact. One example of these controversial initiatives is the 2015 partnership between Royal Free London Trust and Deepmind, an artificial intelligence company owned by Google. As a result of this partnership, NHS patients' health records were shared without explicit consent with Deepmind (Powles & Hodson, 2017). University-based researchers believe that by establishing trust-based relationships with groups of patients and the public, they can help preventing opposition from the public towards the reuse of patient data for research purposes in universities.

Strategies such as integrating citizens' juries and setting up patient groups are good initiatives as they reflect an interest in engaging in meaningful conversations with patients and the public. Whereas transparency and accountability and public participation are key components that reflect good practice (Understanding Patient Data, 2020), the mere implementation of these mechanisms does not guarantee a positive effect. There are risks associated with these initiatives; they can be tokenistic or fail to represent the widest range of social groups whose data is at stake. Thus, it becomes crucial to monitor and evaluate such practices constantly. It could also be argued that while public and patients' involvement in projects could contribute to achieving more 'just' uses of patient data, further actions are needed. For example, it is essential to guarantee responsible data governance to avoid the use of patient data that could lead to the exacerbation of already existing health inequalities and the systematic surveillance and privacy invasion of vulnerable groups (Dencik, Hintz, et al., 2019). Stronger legal protections, reviewed and updated to reflect changing practices, may be required. Considering this, it could be problematic to try to use the public voice and endorsement with the objective of fostering the movement of data for research purposes. The support from the public is an important element, but this should not be the only factor to decide whether data flows or not.

8.5 Infrastructural and data management practices acting as frictions

While advances in technology were recognised as enabling data flows, technical infrastructures and data management practices still are socio-material barriers that block or slow down the

movement of patient data. This has been recognised by other scholars and by the NHS itself (Department of Health, 2012; Keen et al., 2013). There have been some improvements in this context, and the frictions are less than in the past; however, they still are barriers to allowing the data flow. University-based researchers identified two key obstacles: according to them, the quality of data and the infrastructural developments to safely and securely analyse data are not good. This makes it very complicated to work with data. This section discusses these two key barriers.

8.5.1 Actual data work versus the promise of big data

As explained in the first section of this discussion chapter, the key justifications that university-based researchers give for why they are interested in using patient data to conduct research revolve around big data promises. These promises allude to several qualities in large datasets, such as them being accurate representations of the world, and sources that provide answers to previously unsolved questions. However, when university-based researchers and data practitioners talk in detail about their work with data, they highlight the challenges of using datasets initially created to provide care to patients. Whereas the challenges they describe could be used as arguments that contradict big data promises, university-based researchers talked about the promise of big data and the challenges of working with data as if they were completely unrelated matters.

Cleaning data and making sense of data

University-based researchers mainly talked about two key tasks that they have to conduct before reusing data for a specific project. These two activities are "cleaning data" and "making sense of data", and are needed because datasets can have weaknesses. According to university-based researchers, these two tasks are challenging and time-consuming processes that can slow down or block patient data journeys in the very last stage before they can use them in a research project. In other words, cleaning and making sense of data act as data frictions. The fact that university-based researchers and data practitioners talk about the need to clean and make sense of data reveals that they can recognise issues in datasets; however, this does not prompt them to reject the big data promises.

As university-based researchers have explained, once they have managed to overcome other barriers to make data move from the site of data creation (identify the data, request the data, get the data request approved, receive the data), working with patient datasets is not all easy and smooth. According to them, once they receive the data, they have to embark on laborious and

time-consuming labour to get data ready to be used. This endeavour involves several steps, but could probably be summarised if we call it "clean the data and make sense of the data. Cleaning data involves removing "junk data" and deciding what data will be used and what data will be discarded. According to researchers, this is necessary as not all data that arrives in their hands is useful. Making sense of data entails deciphering unclear information and understanding why there are gaps and why certain codes have been assigned to certain types of data.

University-based researchers tended to attribute the weaknesses in datasets to "human actions." For example, they commented that healthcare professionals record data to meet their main objective: delivering the right care to the patient. They also pointed out that they might not have enough time to record all the fields on electronic health records. They also commented that even where standardised codes exist, people will use them in different ways according to their interpretations. This shows that university-based researchers recognise the human intervention in two key moments of the life of data: the initial data production stage and data reuse stage. However it can be argued that despite this recognition of the human intervention, this community does not recognise in their discourses that the data production process is informed by the identity and perspective of those who record data; that it is an error-prone process carrying people's foibles, mistakes and misperceptions (Behar & Gordon, 1995; Keen et al., 2013; Mayer-Schönberger & Cukier, 2013). This also suggests that they overlook that cleaning and making sense of data, which involve making decisions about what variables and attributes will be taken into account and which will be discarded, are inherently subjective processes infused with their biases (boyd & Crawford, 2012). The lack of recognition of the fact that datasets, regardless of their size, are subject to limitations and bias is problematic because if biases and limitations are not understood and outlined, there is a risk of misinterpretation (boyd & Crawford, 2012).

According to several scholars, this human intervention is one of the key reasons why data cannot fall into the description of objective data (Borgman, 2015a; boyd & Crawford, 2012; Kitchin, 2014b). As explained earlier in this work, in the literature of the CDS field, data are commonly understood as material objects constituted through complex socio-material practices (Lupton, 2015). However, even though university-based researchers and data practitioners drew attention to the issues in datasets; this does not defeat the dominant discourse of the promises of data adopted by the university-based researchers' community. On the contrary, what seems to happen is that this discourse obscures data's nature, and obscures that data are

always shaped by their creators. Participants recognise weaknesses in datasets, but these weaknesses seem to disappear when they talk about the big promises of big data.

8.5.2 Expectations concerning the quality of data

Findings of this research show that university-based researchers that have worked with patient data for a considerable time now have accepted that getting what they label as "high-quality data" is extremely difficult. Most importantly, they commented that they are not expecting to have high-quality data and have demonstrated the willingness to work with these datasets even when they are not perfect.

Whereas most of them accept that it is not the obligation of health providers to capture data for secondary purposes accurately, some of them felt shocked or showed discontent with the idea that people are not capturing data. Among the participants, those who show a more empathetic stance have previously worked as healthcare professionals before starting to do research with patient data.

Whereas participants commented that they have accepted that perfect datasets are almost impossible to obtain, they had imagined "the perfect datasets." Ideal datasets, according to them, should be fully populated; this suggests that among university-based researchers and data practitioners, completeness is a very strong marker of quality. They believe that to have ideal datasets, the following is needed: more time for people to record data, redesign of medical codes, and making the completion of electronic records compulsory. University-based researchers have tended to attribute the weaknesses in datasets to people responsible for recording that data in the first place. Or in other words, they believe that data could be an accurate representation of the world if people had the resources and adequate tools to do their job properly. Data quality issues impact data flows, as they can for example slow down the work. This friction can at times frustrate researchers and data practitioners because they imagine that a smoother flow and use of data is possible. Since university-based researchers have embraced the big promises of big data, some of them have engaged in projects aimed at reducing this friction, . However projects of this nature can only be developed if researchers obtain support from their organisations or funders to pursue the promise of frictionless data flows. Therefore, as discussed in Chapter 6, research groups with more access to resources are more likely to engage in these type of projects.

8.5.3 Data safe havens

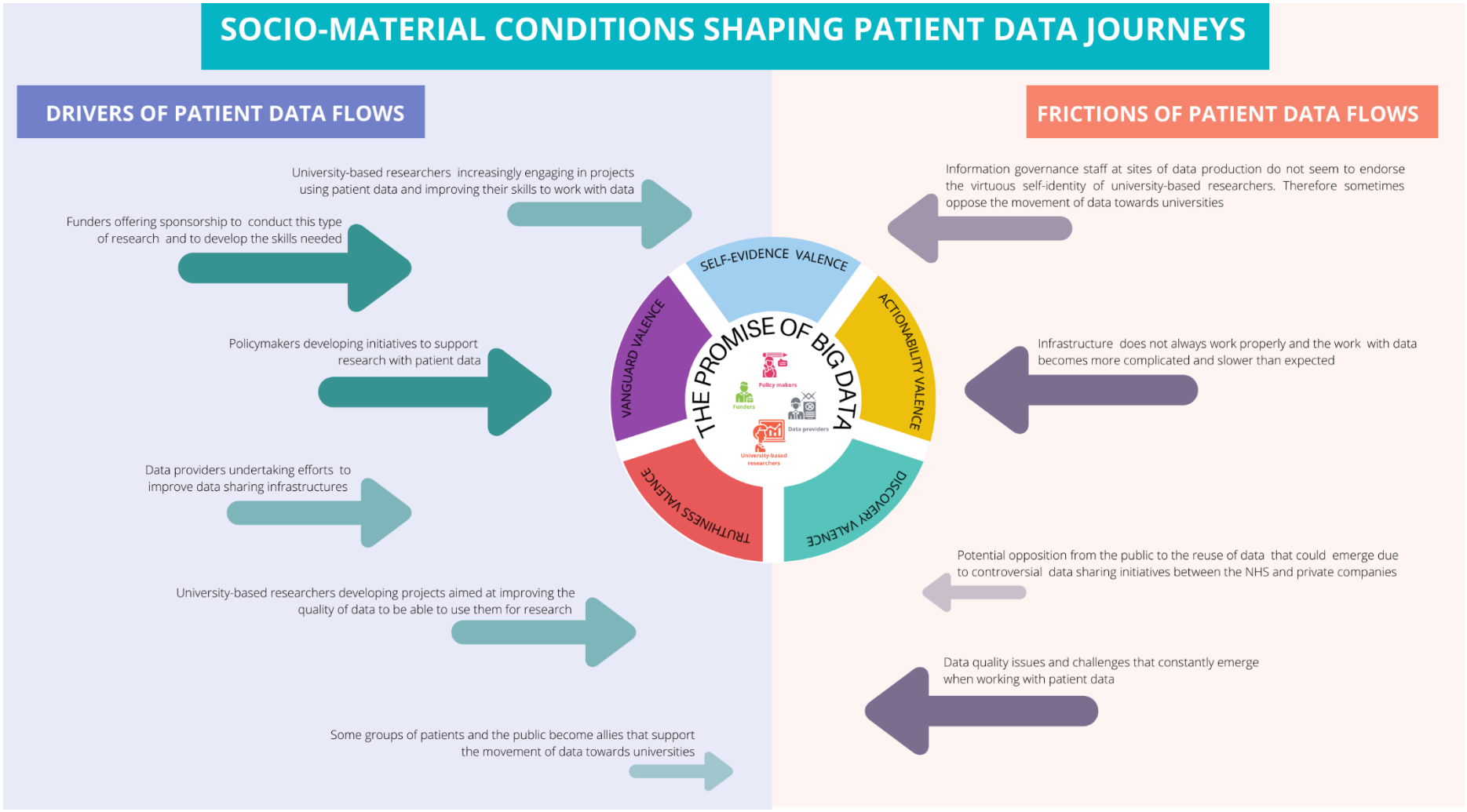
As explained before, data safe havens are secure environments implemented to maintain confidentiality, integrity, and availability of patient data. University-based researchers agree with what many scholars have highlighted in the past, which is that there are several technical difficulties in working in these safe environments.

When talking about their experience using patient data, university-based researchers frequently expressed that working in these environments often has made them feel annoyed or irritated and frustrated as this significantly slows down their work. Despite the technical difficulties of working in these safe environments, university-based researchers seem to accept the use of data safe havens. They have not desisted in their interest in using patient data for research. However, while acknowledging the complications of working within these socio-material environments, they have stressed the need for improving data safe havens. University-based researchers seem to embrace the use of data safe havens for one key reason, and this reason is that they provide a sense of security to data providers, the public, and the university-based researchers' community. Even though data safe havens operate differently and not all of them provide the same level of security, the use of the term “data safe haven” has been useful to reassure data providers, funders and researchers that a project is managing data in the "right way"; that using data safe havens helps to reinforce their trustworthy identity. A factor that perhaps has encouraged university-based researchers to support the use of data safe havens and welcome the idea of investing in improving these environments is that data providers have determined that researchers must store data in a safe haven. Therefore, if they want to be part of the group of people conducting innovative research, they need to comply with this.

8.6 A contrast that is worth highlighting here is the differences in how university-based researchers perceive different types of frictions. On the one hand, university-based researchers have shown opposition to evaluation by external committees and have expressed that they feel that they should be granted access to data based on their good intentions. On the other hand, however, they do not show the same resistance to the imposition of infrastructural barriers. Data safe havens are considered good frictions as they provide a sense of security, and therefore, they embrace them.**Interactions of socio-material factors shaping data flows**

This last section explains and presents a visual representation that illustrates the ways in which sociocultural values and norms interact with existing and emergent material conditions to shape“data journeys” (Bates et al., 2016) of patient data in the UK health sector, with a specific focus on reuse of patient data for research purposes.

Figure 7. Socio-material conditions shaping patient data journeys



The perceptions and expectations of university-based researchers and key stakeholders in the healthcare sector about how data ought to be used appeared to be in harmony. This has helped create a situation in which socio-cultural values and norms interact with existing and emergent material conditions in a way that is favourable to making data flow from the healthcare sector to the hands of university-based researchers. They all have tended to embrace the big promises of big data, which often place great expectations on the results that can be achieved working with large volumes of data and tend to emphasize their benefits sometimes overlooking potential negative implications.

University based researchers are increasingly engaging in research projects reusing patient data. In recent years, it has also been observed that policymakers have developed initiatives to support research using patient data and data providers have undertaken efforts to improve data sharing infrastructures. Furthermore, funding bodies have prioritised offering sponsorship to conduct research with patient data and to develop the skills needed. The actions of these key organisations can motivate and encourage university-based researchers to continue conducting research reusing patient data. This is mainly because this type of research seems to be highly valued, and therefore funding, training, and support are offered to those interested in engaging in health data research. The scenario created by the interaction of these factors has generated a growing demand for data that seemed to be validated and backed up by key funders, data providers and policymakers with access to financial and material resources, visibility and decision making power which has drived the movement of patient data towards universities.

Despite technology advances promoting data flows, technical infrastructures and data management practices are also socio-material frictions that block or slow down the movement of patient data. Infrastructure does not always work properly and the work with data becomes more complicated and slower than expected. Furthermore, data quality and metadata issues constantly emerge and generate challenges when working with patient data which slow down the movement of patient data as university-based researchers are often required to spend long time cleaning and making sense of data. While university-based researchers recognise that these factors can act as barriers, and these could be used as arguments that contradict the big promises of big data, they often talked about the promise of big data and the challenges of working with data as if they were not related. The embrace of the big promises of big data have motivated university-based researchers to engage in projects aimed at improving the quality of datasets and metadata, and solve infrastructural issues that act as barriers for the movement of data towards them. It was observed that barriers generated quality issues in data and metadata

and infrastructure interact with the expectations and assumptions that researchers hold about data. Given that university-based researchers are interested in using patient data because they hold the belief that these type of data have great potential and that conducting research with them is the most innovative way of exploring health issues, they are willing to enage in efforts to foster the movement of data towards universities (eg projects to improve metadata quality). However, as pointed out before, working on these metadata projects involve investing time and financial resources, therefore it is more likely that well-establishead research teams get involved in efforts of this nature.

University-based researchers have developed and projected a virtuous self-identity of their profession. They have tended to present themselves as a group driven by an interest in conducting high-quality and ethical research to benefit patients and the general population. This perception has led university-based researchers to expect smooth flows of patient data towards their groups. The virtuous self-identity of university-based research groups has been endorsed by some funding bodies and data providers, which has driven the movement of patient data towards them. These stakeholders have publicly praised the culture of ethical and excellent research of these groups and expressed their interest in working with them, and opening calls inviting them to make use of patient data produced in the healthcare sector.

Some groups of patients and members of the public have endorsed the virtuous self-identity of university-based researchers. This has helped to some extent to prevent the emergence of frictions of the journeys of patient data towards universities, as this endorsement has prevented opposition to their data practices. While some controversial data sharing initiatives have generated discontent among some members of the public, previous studies have revealed that the public and patients are willing to support reuse of patient data for research purposes if this is conducted in the public benefit and if the adequate measures are in place to ensure the safe and secure handling of data (Skovgaard et al., 2019; Tully et al., 2018). Nonetheless, university-based researchers have tended to perceive that controversial data sharing initiatives have undermined the levels of trust from patients and the public, generating frictions in the circulation of data towards them. Therefore, they have adopted strategies to obtain endorsement from the public and patient groups, often appealing to their virtuous self-identity. University-based researchers have the perception that their initiatives to engage with members of the public and patients have helped them to obtain the support of these groups the support, which has give them the confidence in the legitimacy of their own actions.

The virtuous self-identity of university-based researchers has not always been endorsed by information governance staff at sites of data production and they sometimes have refused sharing data. Thus, the decisions of information governance staff sometimes act as frictions for patient data to flow to university-based research groups. While some researchers have sometimes tried to lubricate the frictions to access patient data generated by the resistance of healthcare institutions to share patient data by presenting the argument that they have the support of patients and the public and by projecting their virtuous self-identity, this has not always been useful. This is because professionals at healthcare organisations do not only consider important the support of these groups, they also expect to be reassured that data is going to be used in a secure way, following ethical principles and guaranteeing that patietns are not going to be negatively impacted. It is worth mentioning that these frictions generated by information governance staff at sites of data production are worth being preserved; it would not be beneficial to open up data flows based on the “good intentions” and expectations of a particular group.However, it was observed that the frustration resulting from the perception that staff at sites of data production are reluctant to share patient data interacts with the frustration around the complexity of regulatory frameworks, physical barriers, and technical challenges to create a sense of significant friction from a variety of social and material sources. It can be argued that it is possible that university-based researchers immersed in their discourse of "virtuous" community, have difficulty in viewing weaknesses and potential negative implications of their data practices. Furthermore, just because something is legal, does not mean that it is ethical and therefore, it should not necessarily be permitted.

Chapter 9. Conclusion

9.1 Overview of the research

The main aim of this thesis has been to gain understanding of how sociocultural values and norms interact with existing and emergent material conditions to shape "data journeys" (Bates et al., 2016) of patient data in the UK health sector, with a specific focus on reuse of patient data for research purposes. It examined where particular types of personal health data flow, and the socio-material drivers and frictions that represent the politics of these data flows. This research also explored the potentials of "Data Journeys" (Bates et al., 2016) as an approach to investigate the movement of personal health data produced in the healthcare sector.

This was achieved by applying Bates et al.'s (2016) Data Journeys methodology with adaptations to fulfil the objectives of this study and utilising the concept of "Data Valences" (Fiore-Gartland & Neff, 2015), which offered a lens to guide the analysis of the different expectations and assumptions about patient data held by university-based researchers.

The main research question posed in this research was: **How do sociocultural values interact with existing and emergent material conditions to generate drivers and frictions that work together to produce the material forms of data journeys in the context of reuse of NHS patients' personal data for research purposes?** In order to answer this question the study aimed to address five objectives.

The first objective was focused on **investigating the different ways in which the movement of data has been studied in the past and exploring the contributions and limitations of such studies**. The literature review revealed that a large number of studies exploring the circulation of data have been conducted; however, often barriers and drivers for the movement of data have been studied with a focus on management, infrastructural or technical issues (Bote & Termens, 2019; De Roo et al., 2016). In the health area, different stakeholders' attitudes towards the reuse of health data for research purposes have been extensively studied. Several studies exploring public and patients' attitudes have been conducted. The perceptions of policymakers and healthcare professionals have also received attention. In this area, some studies have explored researchers' views about the reuse of health data to some extent. However, studies have been mainly concerned with identifying the technical barriers that researchers experience when accessing patient data and how they have worked to overcome these barriers, primarily focusing on data quality issues and anonymisation.

The literature review also showed that in Critical Data Studies (CDS) and related fields, the movement of data has been studied with a different orientation to the studies mentioned above. Studies conducted in CDS and related fields assume that data are socially constructed or socio-material objects, and for this reason cannot be conceived as neutral facts; that the social nature of data produces social effects (e.g. in the case of biased data systems); and that datasets are partial and can never offer complete representations of the world. They also emphasize that power dynamics play a significant role in shaping data flows and that data infrastructures are depictions of knowledge/power that influence the types of questions that can be posed, the ways in which they are addressed, how these answers unfold and who can ask them. A wide range of studies on data circulation have been conducted in CDS and related fields, centring their attention on issues such as the expectations of citizens concerning the uses of data, privacy implications of specific data practices, and the impact that the use of data can have for particular individuals or communities. Nonetheless, little attention in the CDS field has been given to studying health data flows and reuse in the context of the UK's NHS.

The second objective was to **examine selected data journeys of patient data generated in the UK healthcare sector and reused in universities for research purposes**. Through interviews conducted with university-based researchers and experts who are familiar with the use and processing of patient data within the health and care sector (e.g. senior academics, senior members of key organisations such as NHS Digital), and the analysis of documents produced at sites of data reuse explored as part of this study, the researcher mapped selected data journeys of patient data, producing detailed descriptive accounts and diagrammatic representations of these journeys. The journeys of patient data followed were:

Data journey 1: Using technology and data to improve the diagnosis and treatment of stroke. This project was conducted with two key objectives: 1) to develop an overview about the journeys of patients between primary, secondary and community care in Manchester and Salford to gain a better understanding of the patient journey and to identify gaps in the care provided; 2) to propose improvements to support stroke patients and ensure the efficiency and adequate coordination of services.

Data journey 2: Building rapid interventions to reduce antibiotic prescription. This project was conducted with the aim of investigating the drivers for antibiotic prescribing and to study short-term health outcomes after antibiotic prescribing.

Data journey 3: Long-term outcomes of urinary tract infection in childhood (LUCI). This study had the objective of investigating long-term outcomes of children who have experienced a urinary tract infection at any point during their first five years of life.

Data journey 4: Linking electronic health records with passive smartphone activity data to predict outcomes in psychotic disorders. This project was conducted with the aim of investigating clinical outcomes in mental disorders.

Data journey 5: Understanding age inequalities and inequities in the cancer pathway. This project was conducted with the aim of understanding the use of radical rectal cancer treatments, including surgery and neoadjuvant radiotherapy, and their associated outcomes, and evaluating how these varied across England.

The third and fourth objectives were concerned with **gaining insights into the sociocultural values and norms within university-based research groups and the material conditions that together act as factors driving or constraining the movement of patient data towards university-based research groups**. These key insights are summarised below:

1) University-based researchers have tended to embrace the big promises of big data. These assumptions appeared to be in harmony with ideas underpinning the agendas of some data providers, funders and policymakers. This compatibility of perceptions and expectations has contributed to sustaining a growing and validated demand of data which has contributed to driving the flow of data produced in the healthcare sector towards universities to be reused.
2) University-based researchers have strong perceptions about their own profession as well as other actors as reusers of patient data. They have tended to perceive themselves as a virtuous group driven by an interest in conducting high-quality and ethical research to benefit patients and the general population. In contrast, they perceive research groups based at pharmaceutical companies as mainly profit-driven actors and therefore less ethical or unscrupulous. These perceptions have led university-based researchers to expect smooth flows of patient data towards their research groups. Further exploration is needed to understand the way in which this influence data flows.
3) The alignment of views of university-based researchers with agendas of key stakeholders in the heath data landscape has helped to drive the movement of patient data towards universities. In the first instance, university-based researchers have demonstrated a strong interest in seeking opportunities for reusing patient data to

conduct health research. Funding bodies, data providers and policymakers have joined forces to provide the material resources such as funding and infrastructure, which has helped to foster and promote the flow of patient data produced in the healthcare sector to universities.

4) University-based researchers are inclined to believe that controversial data-sharing initiatives between the NHS and external actors have undermined public trust in the type of research they conduct and therefore generated data frictions. This perception has motivated them to develop strategies to gradually build back the public trust. A core element of these strategies is that they have tended to emphasise and project to patients and the public their virtuous self-identity.
5) Despite technological advances, infrastructural barriers and data management practices still slow down or block the movement of patient data from the healthcare sector to universities conducting health research. Whereas university-based researchers recognise that these factors can act as barriers, this does not discourage them from embracing the big promises of big data. Rather, they have shown a great interest in engaging in projects aimed at improving the quality of datasets and solving infrastructural issues that act as barriers for the movement of data towards them, so the big promises of big data can be fulfilled.

The fifth objective of this research was to reflect on **the potential challenges that can be encountered when applying the Data Journeys approach to explore patient data flows and how this approach might be adapted for future use in similar contexts**. This study identified that a key challenge to applying the Data Journeys methodology is a lack of transparency of key stakeholders; this was useful for recognising three key aspects of this methodology not previously articulated. The first one was that the full potential of it can only be realised when stakeholders involved show willingness to participate and are transparent. The second is that the challenges in gaining research access to key sites of practice due to lack of transparency are useful for identifying black boxes of data practices that require the design and application of additional data collection strategies to grey them out. The third is that if this methodology is to be used in similar contexts, a number of adaptations might be useful. For example, where black boxes are identified, it might be necessary to: 1) attend public events to capture the public discourses of people at this sites and key aspects of their culture; 2) invest significant time to try and build contacts with key informants at the identified black boxes to try

to develop an understanding of what is happening at these sites over time; 3) draw on alternative sources to collect data, for example journalistic reports or Freedom of Information requests.

9.2 Contribution to knowledge

The contributions of this research could be of interest to academic scholars from different fields, such as information studies, health studies, and critical data studies. As follows, the empirical, theoretical and methodological contributions of this thesis are presented.

9.2.1 Empirical contribution

This study draws attention to the critical role that sociocultural values and norms and material factors play in shaping data movement. It offers insights into the motivations that researchers have to seek access and conduct research with large patient datasets. In addition to this, this study confirms that, as identified in other studies exploring the circulation of data, despite technological advances, infrastructural barriers still slow down or block the movement of patient data from the healthcare sector to universities for conducting health research. The study recognises the value of studies previously conducted in the Information Studies field on the circulation of data with a techno-managerial approach. Indeed, many participants drew attention to the existence of such management, infrastructural or technical issues that act as barriers for the movement of data. However, through interpreting university-based researchers' perceptions through a CDS lens we are able to generate novel empirical insights.

As observed by scholars in the Critical Data Studies field, detailed empirical research is needed to make sense of data alongside conceptual and philosophical reflection. Several scholars have contributed to developing a better conceptual and theoretical understanding of data. However, while there are some in-depth studies in this field, including Aula's work focused on health data flows in the Finnish healthcare sector, more work in this area is needed. Back in 2014, Kitchin highlighted the urgent need to conduct detailed empirical research to advance the understanding "of both the overall construction of data assemblages and their apparatus and individual elements" (Kitchin, 2014c, p. 6). This in-depth study responded to this call by exploring the sociocultural values and norms within university-based research teams (e.g. their motivations for data practices, patterns of decision-making, internal dynamics, perceptions of their own culture, and ways in which they relate and engage with external cultures), and how they interacted with material elements to generate the socio-material conditions that create

drivers and frictions that work together to produce the material forms of data journeys in the context of reuse of NHS patients' data for research purposes.

This research contributes to a wider debate within the CDS field about how different social groups have adopted promises of big data and how this has impacted flows of data (boyd & Crawford, 2012; Fiore-Gartland & Neff, 2015). For example, the study observes that university-based researchers' discourses appear to be aligned with key stakeholders' discourse in the healthcare landscape, such as some data providers (e.g. NHS Digital), funders and policymakers. This alignment has been helpful to make data flow from the healthcare sector to university-based researchers as this has fostered the creation and validation of a growing demand for patient data. This work also observed that, motivated by their embrace of big data promises, some researchers have engaged in projects to remove barriers generated by infrastructure and data management practices. They have done this with the expectation that they and other university-based research teams can have access to patient data with fewer constraints.

The study also drew attention to the important role that self-identities can play in shaping data circulation. It observed that the self-identification of university-based researchers and data practitioners as a virtuous community has contributed to shaping university-based researchers' expectations regarding data access and has influenced how they engage with other actors. This self-perception has been helpful as some key funders and data creators have been inclined to endorse this identity and accordingly facilitate access to patient data. Besides, this study observed that university-based researchers have tended to perceive pharmaceutical-based research teams as primarily profit-driven, and in consequence, less deserving of access to patient data.

This work contributes to a wider academic discussion in the literature about the Care.data scandal's aftermath. A number of studies have been conducted to understand the attitudes of different stakeholders concerning this initiative and the factors that led to its failure. Whereas other studies had pointed out that no systematic differences in attitudes towards data were identified among members of the public before and after the Care.data case disclosure, this research observed that university-based researchers believe the opposite. The study reveals that university-based researchers have the perception that the Care.data scandal significantly damaged the trust of the public in data sharing initiatives between the healthcare sector and universities. It also sheds light on the efforts that university-based researchers and data practitioners have undertaken intending to rebuild the trust that they perceived as lost.

9.3 Theoretical contribution

Integration of Data Journeys and Data Valences

This research brought together two different strands of thinking to study the movement of patient data: Data Journeys and Data Valences. Data journeys were explored and a reflection on the findings from the fieldwork was conducted using the Data Valences lenses. Bringing together Data Journeys and Data Valences is an original contribution because this is the first study that combines these two different approaches to explore the movement of data.

The Data Journeys methodology was applied systematically due to its usefulness in exploring where particular types of personal health data flow, and the sociocultural values and norms interacting with existing and emergent material conditions that represent the politics of these data flows. The understanding gained by applying the Data Journeys approach was complemented by utilising Data Valences, a conceptualisation proposed by Fiore-Gartland & Neff (2015). As explained in previous chapters, this concept refers to the broad range of expectations and values that are expressed in people's discourses and practices. The notion of Data Valences can be particularly helpful to examine how data are valued and interpreted and what is expected from them across different contexts, and to develop an understanding of how data work and perform in different social settings.

Through the integration of Data Journeys and Data Valences, it was possible to develop a diagrammatic representation (see page 190) that illustrates how different socio-material factors work together in shaping the movement of patient data generated in the UK healthcare sector. The diagram does two main things. In first instance, it integrates five data valences (self-evidence, truthiness, discovery, actionability and vanguard) that highlight the key expectations and assumptions about data expressed in the discourses and actions of university-based researchers and key stakeholders in the healthcare landscape, such as some funders, data providers and policymakers; and how these valences combined show that these actors have tended to embrace the big promises of big data. Second, it shows how despite these big promises of big data that often placing great expectations oand tending to emphasize the benefits of working with large volumes of data, there are drivers and frictions that shape the journeys of patient data.

The proposition of an additional Data Valence

Fiore-Gartland and Neff identified six different data valences; these are: 1) self-evidence, 2) truthiness, 3) discovery, 4) actionability, 5) transparency, and 6) connection. This study's

second theoretical contribution is the addition of the **vanguard** valence, a data valence distinct to those outlined by Fiore-Gartland & Neff (2015). This valence was observed in the discourses of university-based researchers. People evoked the vanguard valence when they talked about doing research with large patient data sets as a practice that positions them as part of a group breaking with old ways of doing science, as innovative researchers willing to introduce and explore new ways of doing health research.

9.4 Methodological contribution

The creators of Data Journeys piloted the methodology in an experimental and exploratory way, looking at meteorological and climate data flows. As discussed in the literature review, since its creation, this methodology has influenced a number of studies; however, this is the first study that systematically applied the Data Journeys methodology to explore the journeys of patient data created in the UK healthcare sector. The systematic application of Data Journeys allowed me to recognise three new and key aspects of this method as articulated in response to objective 5 above (see page 196).

9.5 Recommendations for university-based researchers

This study helps to understandthe culture of of university-based researchers who work with patient data, how it relates to other cultures, and how that impacts the flow of patient data generated in the UK healthcare sector. Drawing from the findings of this work, this section presents recommendations that could be helpful for university-based researchers.

Findings of this study suggest that university-based researchers have tended to embrace the big promises of big data, which often place great expectations on the results that can be achieved working with large volumes of data and tend to pay attention to the potential benefits of reusing this data while overlooking potential negative implications. As pointed out by participants, valuable results can be achieved through research conducted with patient data; however, a more nuanced approach is needed. In light of this, it would be valuable to invite university-based researchers who work with patient data to pay greater attention to the fact that datasets can never offer objective or accurate representations of the health of the country because in reality complex and often hidden social and political processes influence their creation. Likewise, it is important to always consider that more data is not necessarily equal to better insights because such insights can only be achieved after those volumes of data have been cleaned, analysed and interpreted. These processes not only require human intervention, making them subject to

biases, but also demand time and material resources. Equally important is to not forget that finding patterns in data is not equivalent to having a better understanding of health issues at population scales, and that knowing the answers to research questions on its own does not lead to addressing health issues on the ground. Rather, such answers can inform practice and decision-making processes, which could potentially lead to useful results such as the design of better health interventions or policy for the public good.

Findings of this study suggest that university-based researchers who work with patient data perceive themselves as a virtuous group of people doing health research motivated by an interest in providing benefits for the UK population. This self-perception has been used as a justification to data providers for expecting smooth flows of data towards them and is also embedded at the core of engagement initiatives advocating for the public and patients' support for health data research. This research has shed light on the important role that this virtuous self-identity plays in shaping the flows of patient data and invites researchers to recognise themselves as political actors and adopt a reflective stance towards their own practices and discourses.

9.6 Future research

This study has advanced our understanding of the ways in which sociocultural values and norms interact with existing and emergent material conditions to shape the movement of patient data. Nonetheless, the scope could be expanded building on what was done in this research, or alternatively, different methods could be used to gain new insights. The insights gained could be expanded by travelling to more sites of data production, processing and reuse along the data journeys to conduct additional interviews with other stakeholders such healthcare practitioners, policymakers, funding bodies' representatives, information governance staff at sites of data production, patients and the public to get a deeper understanding of the tensions and frictions as well as the drivers that shape the movement of data. Additionally, field observations at different sites of data practice could be carried out; this is one of the data collection strategies suggested by the Data Journeys approach not used in this research that has the potential to enrich the insights into the culture within university-based research groups. A long term strategy could also be developed to build contacts with key informants in some of the black boxes identified in this study (e.g. pharmaceutical companies, data intermediary sites) to try and build up a picture over time about what is happening in some of these spaces.

This study centred its attention on the UK context. Future research could explore the socio-material factors that shape the movement of patient data in different countries; this could offer interesting and novel comparative insights about how different sociocultural and material factors interact and shape the socio-material conditions for the circulation of health data in contexts different to the UK.

Appendix 1. Two different uses of the data journeys metaphor

Bates et al. (2016) and Leonelli (2016) used the metaphor of data journeys in their work, however each of them with a different interpretation, focus and disciplinary approach. To avoid confusion, the key characteristics of these two approaches are summarised in the table below.

	Bates et al. (2016)	**Leonelli (2016)**
Disciplinary approach	Cross-disciplinary approach, closely related to the CDS approach, which is characterised by paying attention to power, political, ethical, and social issues related to data (Iliadis & Russo, 2016c).	Developed with a Science and Technology Studies approach (STS). work provides a
Key focus	Concerned with the movement of data by taking into account the broader context in which data movements take place.	Interested in conducting a philosophical analysis of data handling practices characterised by presenting thick and detailed descriptions and explanations of processes related to these practices.
Main points of interest	-Interested in the cultural values, power dynamics, and institutional and political contexts that shape data flows across different sites of practice and sectors; as well as in potential, blocked and lack of data movement, the speed and timing of these movements and intersecting journeys resulting from the combination of data from different sources (Bates et al., 2016). -Pays attention the ways in which data journeys connects social actors in different and new ways across different sites and places.	-Interested in the technical conditions that are necessary for data to move in different research situations, specifically in the field of biology, and in the diverse ways in which the value of data can be determined. -Strong emphasis in techniques and procedures involved in the handling of data (Leonelli, 2016). -Pays attention to processes of packaging data that 'prepare data to travel' (e.g. the classification, selection, formatting and standardisation of data) and the creation and utilisation of methods to analyse, retrieve, visualise and control the quality of data (Leonelli, 2016).
Use of the term data journey	1. As a concept according Bates et al (2016), data journeys refer to the movement of data across space and time, through different cultures and sites of practice from their initial generation through to reuse in diverse contexts. 2. Data journeys is also utilised by Bates et al (2016, p. 1) to refer the methodology these authors developed with the objective of exploring the political, social and cultural factors which shape the movement of data; as they travel between different cultures and sites of data practice, initiating from their generation through to reuse in a number of diverse contexts.	Leonelli (2016, p. 39) utilises data journeys as a concept to refer to the movement of scientific data from the initial site in which they are generated, to different sites within or beyond the research field in which they are produced.

Appendix 2. Interview guide – Phase 1

Participants: In phase 1 of the research, we are inviting experts who are familiar with the use and processing of personal patient data within the health and care sector.

Style of interview: conversational

Duration: 30-45 minutes. This interview is divided in two parts.

First part:

The researcher will ask you to talk about your professional background and current position.

Second part:

You will be asked to talk about examples of data flows in the healthcare sector in which from your perspective:

- patient data flows too easily between different people and/or organisations.
- there are too many barriers restricting the movement of patient data between different people or organisations.

When addressing these topics, you might want to consider: type of data that is collected, type of data that is shared, type of data that is not shared, people involved in these data flows.

Main challenges to overcome when sharing patient data for purposes other than the direct care of patients.

The role of existing relevant guidance and regulations within data flows in the healthcare sector.

This research has been approved by The University of Sheffield Research Ethics Committee.

* Interviews will be recorded using a digital recording device and field notes will be taken.

* Data will be anonymised and computer files will be coded with a random number, unless you wish to be identified.

* The results of this study will be included in my PhD thesis which will be publicly available.

Appendix 3. Interview guide – Phase 2

About the role of the interviewee and the organisation

1. What is your background (academic/professional) and area of expertise?
2. What is your role and key responsibilities and what attracted you to this position?
3. For how long have you been working in this organisation and in this team?
   - When you joined, did you start working in a different role?
   - How does the organisation has changed over the years you have been here?
   - What do you think are the reasons for these changes?
4. Can you tell me a little about your team? (e.g. structure, team members, main responsibilities)
5. How would you describe the culture in this project/team?
6. How do you think it fits with the wider culture of this organisation?

The work with data

1. What type of patient data do you collect?/What type of patient data do you use?
2. What do you do with the data when data gets to this department? - Can you tell me about a good/interesting example? Can you think in another different/contrasting example?
   - Who was involved? What did you do with these data?
   - How did you make sense of and interpret these data?
   - In this example- did you evaluate the quality of data and how do you know/do you assess if you can trust the data you are working with?
3. How does the way in which you get access to/process patient data has changed over the years?
   - Can you tell me about an example of this? AND How do you feel about these changes?

Sharing data for secondary uses

1. To whom is the data processed here made available and how is access to these data obtained?
   - How do you feel about making patient data available to different organisations such as charities, pharmaceutical companies, research units? How do you relate to those differently?
2. Why do you feel is important to share these data for secondary purposes?. .and who benefits if that happens?
   - Do you feel that some secondary uses of data should be prioritized over others?
   - Do you feel that some current uses of patient data should be restricted?
3. There are a number of regulations that dictate how and under which circumstances data can be shared and to whom. What is your opinion in relation to these regulations? Do you think they are too restrictive? Do you think they could be improved in some way?
4. The organisation in which you work, collects patient data from other sites, then processes that data and shared it with a number of organisations/projects. Do you feel that these different sites share some “organisational values” with this organisation?

    - How would you describe the culture of those sites?
5. What kind of data requests do you receive and how are these requests evaluated? (e.g. what is the criteria for approving or rejecting them)
    - Can you tell me about a couple of recent examples of this?
    - Can you tell me about a recent example where you have rejected a data request?
6. How do you feel about the ways in which different organisations use patient data?
7. What do you think is driving to making efforts for making data available to third parties? –and how do you feel about that?
8. In your opinion are data well documented when they arrive to this organisation?

Appendix 4. Information sheet – Phase 1

Socio-material factors shaping patient data journeys in the United Kingdom

Researcher: Itzelle Medina Perea, The University of Sheffield,

Supervisory team: Dr Jo Bates, Dr Andrew Cox

This study will centre its attention in the movement of patient data within the healthcare sector. While it is true that data flows or the movement of data through space has been explored within the Information Studies field, many of the studies have only centred its attention on the internal dynamics of information systems and knowledge infrastructures without paying attention to the broader socio-material contexts and power dynamics that influence their development. Data flows in the health care sector have been examined, primarily from an information management perspective. The study proposed here, is distinct from these previous studies as it does not intend to take an information management approach to investigating the flow of data; rather, it pays attention to sociological aspects of the movement of data.

The aim of this research is to gain understanding of the socio-cultural factors that influence the movement of patients´ personal data within the health care sector from the initial generation of data through to use for secondary purposes in diverse contexts. It will examine where particular types of personal health data flow, what are the socio-cultural drivers and restrictions of these data flows, and who is involved in the process of shaping these flows.

Research project structure

Phase 1: Identification of potential data journeys to follow

The first stage of the research is exploratory in nature; it has the objective of defining what data journeys to be followed. In this phase the researcher will conduct interviews with experts who are familiar with the use and processing of patient data within the healthcare sector.

Additionally, in this initial stage a documentary analysis will be conducted to developing understanding of key data sharing policies and legislation. The insights provided by these experts and the documentary analysis will be useful to define potential journeys to follow.

Phase 2: Following data journeys

The second stage of this research consists in defining the data flows to follow and embarking in the process of following data journeys. The study proposed here will focus on two contrasting data journeys, one where there are perceptions that data flows too freely, and one where there are perceptions that data do not flow freely enough.

Appendix 5. Consent form – Phase 1

The University of Sheffield Information School	The socio-cultural life of personal health data flows in the UK healthcare sector

Researcher

Name: Itzelle Aurora Medina Perea Contact information: iamedinaperea1@sheffield.ac.uk
(Supervisors: Dr Jo Bates, Dr Andrew Cox.)

Purpose of the research

The aim of this research is to gain understanding of the socio-cultural factors that influence the movement of patients´ personal data within and external to the health care sector in the UK. It will examine where particular types of personal health data flow and the socio-cultural drivers and restrictions shaping these data flows. The study will focus on two contrasting data journeys - one where there are perceptions that data flows too freely, and one where there are perceptions that data do not flow freely enough.

Research objectives
In phase 1 of the research (to which you are being invited to participate), we aim to identify which data journeys to focus on in the main study, by exploring with expert practitioners their experiences of data sharing and movement of personal patient data produced within the NHS.

Who will be participating?

In phase 1 of the research, we are inviting experts who are familiar with the use and processing of personal patient data within the health and care sector.

What will you be asked to do?

Participants will be asked to participate in a 30-minute interview about the use and processing of data of patients within the health and care sector. The interviews will be conducted by the researcher face to face or via Skype, at your convenience.
We will ask you questions about e.g.

- Can you think of any good examples of cases where you think personal patient data flows too easily between different people/organisations, or cases where you think there are too many barriers restricting the movement of data between different people/organisations? What do you think the reasons are for these challenges?
- What do you perceive to be the most challenging barriers to overcome when sharing patient data for purposes other than the direct care of patients?

What are the potential risks of participating?

The risks of participating are the same as those experienced in everyday life.

What data will we collect?

Interviews will be recorded using a digital recording device and field notes will be taken.

What will we do with the data?

I will transcribe the audio recording into text for analysis. I will be analysing the data collected for inclusion in my PhD thesis. The data will be stored on the University of Sheffield Information School's research data drive which can be accessed by only by me, my supervisors, and the School's Examinations Officer and ICT staff operating the facility. I will also store a password protected back up copy on my personal computer.

Will my participation be confidential?

Data will be anonymised and computer files will be coded with a random number, unless you wish to be identified.

What will happen to the results of the research project?

The results of this study will be included in my PhD thesis which will be publicly available. Please contact the School six months after 01/03/2021. We also plan to report the findings of this research in peer-reviewed journals and conferences.

- I confirm that I have read and understand the description of the research project, and that I have had an opportunity to ask questions about the project.
- I understand that my participation is voluntary and that I am free to withdraw at any time without any negative consequences.
- I understand that if I withdraw I can request for the data I have already provided to be deleted, however this might not be possible if the data has already been anonymised or findings published.
- I understand that I may decline to answer any particular question or questions, or to do any of the activities.
- I understand that my responses will be kept strictly confidential, that my name or identity will not be linked to any research materials, and that I will not be identified or identifiable in any report or reports that result from the research, unless I have agreed otherwise.
- I give permission for all the research team members to have access to my responses.
- I give permission for the research team to re-use my data for future research as specified above.
- I agree to take part in the research project as described above.

Participant Name (Please print)	Participant Signature
Researcher Name (Please print)	Researcher Signature

Date

Note: If you have any difficulties with, or wish to voice concern about, any aspect of your participation in this study, please contact Dr Angela Lin, Deputy Research Ethics Coordinator, Information School, The University of Sheffield (ischool_ethics@sheffield.ac.uk).

Appendix 6. Information sheet – Phase 2

The University of Sheffield Information School	The socio-cultural life of personal health data flows in the UK healthcare sector INFORMATION FOR PARTICIPANTS

Researcher

Name: Itzelle Aurora Medina Perea Contact information: iamedinaperea1@sheffield.ac.uk
(Supervisors: Dr Jo Bates, Dr Andrew Cox.)

Purpose of the research

This study explores socio-cultural dynamics shaping the journey of selected types of personal patient data produced within the NHS. It examines where specific types of patient data flow, who is involved in the process of shaping these flows, and the socio-cultural drivers and restrictions of these data flows.

Expected outcomes:

- Generate new knowledge about the ways in which socio-cultural factors influence the movement of patient data in the current era, and consider what this means for how data flows bring patients into different forms of relation with other social actors.
- Develop recommendations for implementing "just" practices in data sharing, and make patient data flows more clear and transparent.

Research project structure

Through following the journey of specific types of patient data between different sites of practice, the researcher intends to gain understanding about how socio-cultural values and practices are connected in the transformation and movement of data on its journey from production to re-use.

The study consists of two phases. The objective of phase 1 (now completed) is to identify a number of potential data journeys to explore in the second phase. The second phase (in which you are invited to participate) consists of further refining the specific data journeys to follow and embarking on the process of following the data as they journey through different organisations, projects etc.

Data collection:

Phase 1: Identification of potential data journeys to follow (completed)

In phase 1, the researcher conducted interviews with experts who are familiar with the use and processing of patient data within the healthcare sector. Additionally, a documentary analysis was conducted to develop deeper understanding of key data sharing policies and legislation. Drawing from the findings of phase 1, it has been decided that this study will focus on journeys of data that: a) are collected in the primary care and shared with different organisations for uses other than the direct care of patients, or b) are collected via self-monitoring and digital healthcare-related wearable devices provided by commercial companies to NHS patients.

Phase 2: Following data journeys (in which you are invited to participate)

The researcher will conduct interviews with people who work in different organisations or projects that use patient data, and (if consent is obtained) observe what people do in their work and the environment that they work in. The researcher will also conduct documentary analysis of policy, legislation and other relevant documents relevant to the movement of data through that site.

Who will be participating?

In phase 2 of the research, we are inviting team members of different projects or organisations who are familiar with the use and processing of personal patient data.

What will you be asked to do?

Participants will be asked to participate in a 45-minute interview about the use and processing of data about patients in their organisation/project.

What are the potential risks of participating?

The risks of participating are the same as those experienced in everyday life.

What data will I collect?

Interviews will be audio recorded using a digital recording device and field notes will be taken.

What will I do with the data?

I will transcribe the audio recording into text for analysis. I will be analysing the data collected for inclusion in my PhD thesis. The data will be encrypted and stored on the University of Sheffield Information School's research data drive which can be accessed by only by me, my supervisors, and the School's Examinations Officer and ICT staff operating the facility. I will also store an encrypted password protected back up copy on my personal laptop.

Audio-recordings may be sent to a transcription service agency which will be required to sign a non-disclosure agreement to ensure confidentiality of data is protected.

Data will be retained by the researcher and supervisors to use it on future research projects. The data will be stored in the Information School Research data drive and in my personal laptop.

Will my participation be confidential?

Data will be anonymised and computer files will be coded with a number, unless you wish to be identified.

If you have a unique job role, or are one of only a few people working on a particular project, it may be possible for knowledgeable individuals to guess who you are even if your quotations are anonymised. If this is a concern for you please speak to the researcher, so that I can take appropriate measures to anonymise the data when I write up my findings.

What will happen to the results of the research project?

The results of this study will be included in my PhD thesis which will be publicly available. I also plan to report the findings of this research in peer-reviewed journals and conferences.

What is the legal basis for processing your personal data?

The University of Sheffield will act as the Data Controller for this study. This means that the University is responsible for looking after your information and using it properly. In order to collect and use your personal information as part of this research project, we must have a basis in law to do so. The basis that we are using is that the research is 'a task in the public interest'.

Appendix 7. Consent form– Phase 2

The University of Sheffield Information School	The socio-cultural life of personal health data flows in the UK healthcare sector

Researcher

Name: Itzelle Aurora Medina Perea Contact information: iamedinaperea1@sheffield.ac.uk
(Supervisors: Dr Jo Bates, Dr Andrew Cox.)

Purpose of the research

This research examines the socio-cultural dynamics that promote, slow down, or block the movement of particular forms of personal patient data between different social actors within, and beyond, the health care sector in the UK. This study will apply the Data Journeys approach. This approach is useful to explore the social, cultural, and political factors shaping the movement of data between different sites and cultures of data practice.

Expected outcomes:

- Generate new knowledge about the ways in which socio-cultural factors influence the movement of patient data in the current era, and consider what this means for how data flows bring patients into different forms of relation with other social actors.
- Develop recommendations for implementing "just" practices in data sharing, and make patient data flows more clear and transparent.

Research project structure

This research project has two main phases, and these phases are explained as followed:

Phase 1: Identification of potential data journeys to follow (completed)
The first stage of the research was exploratory in nature; it has the objective of defining what data journeys to follow in the main phase of the research. In phase 1, the researcher conducted interviews with experts who are familiar with the use and processing of patient data within the healthcare sector. Additionally, a documentary analysis was conducted to develop deeper understanding of key data sharing policies and legislation. Drawing from the findings of phase 1, it has been decided that this study will focus on journeys of data that: a) are collected in the primary care and shared with different organisations for uses other than the direct care of patients, or b) are collected via self-monitoring and digital healthcare-related wearable devices provided by commercial companies to NHS patients.

Phase 2: Following data journeys
The second stage of this research (in which you are invited to participate), consists of further refining the specific data journeys to follow and embarking on the process of following the data as they journey through different organisations, projects etc. At each of these organisations, projects etc (sites of data practice) the researcher will visit for 1-2 days to conduct interviews with people who work there and undertaken field observations aimed at getting a better understanding of the culture of the team/project/organisation. The researcher will also conduct documentary analysis of policy, legislation and other relevant documents relevant to the movement of data through that site. Key informants will be identified through desk research and through snowball sampling techniques that will be adopted once in the field.

Who will be participating?

In phase 2 of the research, we are inviting practitioners working in organisations and projects relevant to the selected data journeys, and who are familiar with the use and processing of personal patient data.

What will you be asked to do?

Participants will be asked to participate in a 45-60-minute interview about the use and processing of data about patients related to their work. The interviews will be conducted by the researcher face to face or via Skype.

What are the potential risks of participating?

The risks of participating are the same as those experienced in everyday life.

What data will I collect?

Interviews will be audio recorded using a digital recording device and field notes will be taken.

What will I do with the data?

I will transcribe the audio recording into text for analysis. I will be analysing the data collected for inclusion in my PhD thesis. The data will be encrypted and stored on the University of Sheffield Information School's research data drive which can be accessed by only by me, my supervisors, and the School's Examinations Officer and ICT staff operating the facility. I will also store an encrypted password protected back up copy on my personal laptop.

Audio-recordings may be sent to a transcription service agency which will be required to sign a non-disclosure agreement to ensure confidentiality of data is protected.

Data will be retained by the researcher and supervisors to use it on future research projects. The data will be stored in the Information School Research data drive and in my personal laptop.

Will my participation be confidential?

Data will be anonymised and computer files will be coded with a number, unless you wish to be identified.

If you have a unique job role, or are one of only a few people working on a particular project, it may be possible for knowledgeable individuals to guess who you are even if your quotations are anonymised. If this is a concern for you please speak to the researcher, so that I can take appropriate measures to anonymise the data when I write up my findings.

What will happen to the results of the research project?

The results of this study will be included in my PhD thesis which will be publicly available. Please contact the School six months after 01/03/2021. I also plan to report the findings of this research in peer-reviewed journals and conferences.

What is the legal basis for processing your personal data?

The University of Sheffield will act as the Data Controller for this study. This means that the University is responsible for looking after your information and using it properly. In order to collect and use your personal information as part of this research project, we must have a basis in law to do so. The basis that we are using is that the research is 'a task in the public interest'.

Declaration of consent

- I confirm that I have read and understand the description of the research project, and that I have had an opportunity to ask questions about the project.
- I understand that my participation is voluntary and that I am free to withdraw at anytime without any negative consequences.
- I understand that if I withdraw I can request for the data I have already provided to be deleted, however this might not be possible if the data has already been anonymised or findings published.
- I understand that I may decline to answer any particular question or questions, or to do any of the activities.
- I understand that my responses will be kept strictly confidential, that my name or identity will not be linked to any research materials, and that I will not be identified or identifiable in any report or reports that result from the research, unless I have agreed otherwise.
- I give permission for all the research team members to have access to my responses.
- I give permission for the research team to re-use my data for future research as specified above.
- I agree to take part in the research project as described above.

Participant Name (Please print)	Participant Signature
Researcher Name (Please print)	Researcher Signature

Date_____

Note: Further information, including details about how and why the University processes your personal information, how we keep your information secure, and your legal rights (including how to complain if you feel that your personal information has not been handled correctly), can be found in the University's Privacy Notice https://www.sheffield.ac.uk/govern/data-protection/privacy/general.
If you have any difficulties with, or wish to voice concern about, any aspect of your participation in this study, please contact Dr Angela Lin, Deputy Research Ethics Coordinator, Information School, The University of Sheffield (ischool_ethics@sheffield.ac.uk).

Appendix 8. Research ethics approval letter – Phase 1

The University Of Sheffield.

Downloaded: 26/02/2021
Approved: 25/04/2018

Itzelle Medina Perea
Registration number: 160264304
Information School
Programme: INFR33 PhD

Dear Itzelle

PROJECT TITLE: The socio-cultural life of personal health data flows in the UK healthcare sector
APPLICATION: Reference Number 018069

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 25/04/2018 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 018069 (form submission date: 17/04/2018); (expected project end date: 01/02/2020).
- Participant information sheet 1040221 version 1 (22/02/2018).
- Participant consent form 1040222 version 3 (17/04/2018).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Daniel Rose
Ethics Administrator
Faculty of Social Sciences

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/approval-procedure
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

Appendix 9. Research ethics approval letter– Phase 2

The University Of Sheffield.

Downloaded: 26/02/2021
Approved: 29/08/2018

Itzelle Medina Perea
Registration number: 160264304
Information School
Programme: INFR33 PhD

Dear Itzelle

PROJECT TITLE: The socio-cultural life of personal health data flows in the UK healthcare sector
APPLICATION: Reference Number 022551

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 29/08/2018 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 022551 (form submission date: 07/08/2018); (expected project end date: 01/02/2020).
- Participant information sheet 1050213 version 4 (07/08/2018).
- Participant information sheet 1050529 version 3 (07/08/2018).
- Participant consent form 1050214 version 1 (25/07/2018).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Daniel Rose
Ethics Administrator
Faculty of Social Sciences

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/rs/ethicsandintegrity/ethicspolicy/approval-procedure
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

Abbreviations

ALF	Anonymised Linking Field
BCI	Bowel Cancer Intelligence
BRIT	Building Rapid Interventions to Reduce Antibiotic Resistance
CHC	Connected Health Cities
CORECT-R	Colorectal Repository
CPRD	Clinical Practice Research Datalink
CRIS	Clinical Record Interactive Search
DOB	Date Of Birth
DSA	Data Sharing Agreement
DUTY	Diagnosis of Urinary Tract infection in Young children
EHR	Electronic Healthcare Record
EURICA	Epidemiology of Urinary Tract Infection (UTI) in Children with Acute Illness in Primary Care
GDPR	General Data Protection Regulation
GP	General Practice
GRO	General Register Office
HASU	Hyper Acute Stroke Unit
HES	Hospital Episode Statistics
HRA CAG	Health Research Authority's Confidential Advisory Group
IGARD	Independent Group Advising on the Release of Data
IGRP	Information Governance Review Panel
IMD	Index of Multiple Deprivation
LIDA	Leeds Institute for Data Analytics
LUCI	Long-term outcomes of urinary tract infection in childhood
MCCD	Medical Certificate of Cause of Death
NCRAS	National Cancer Registration and Analysis
NHAIS	National Health Application and Infrastructure Services
NHS	National Health Service
NWAS	North West Ambulance Service

NWIS	NHS Wales Informatics Service
ODR	Office for Data Release
ONS	Office for National Statistics
OPD	Outpatient Data
PEDW	Patient Episode Database for Wales
PHE	Public Health England
PROMS	Patient Reported Outcome Measures Survey
PROMS	Patient Reported Outcome Measures Survey
SAIL	Secure Anonymised Information Linkage databank
SLAM	South London and Maudsley NHS Foundation Trust
SLAM BCR	South London and Maudsley Biomedical Research Centre
SRFT	Salford Royal NHS foundation Trust
SUS	Secondary Uses Service
TRE	Trustworthy Research Environment
UTI	Urinary Tract Infection
WECC	Welsh Electronic Cohort of Children

Bibliography

Aitken, M., de St Jorre, J., Pagliari, C., Jepson, R., & Cunningham-Burley, S. (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. https://doi.org/10.1186/s12910-016-0153-x

Annan-Callcott, G. (2021, September 2). *What we know about communicating how health data is used | Understanding patient data*. Understanding Patient Data. http://understandingpatientdata.org.uk/news/what-we-know-about-communicating-how-health-data-used

Asthma UK. (2018). *Data sharing and technology: Exploring the attitudes of people with asthma*. https://www.asthma.org.uk/globalassets/get-involved/external-affairs-campaigns/publications/data-report/data-sharing-and-technology---exploring-the-attitudes-of-people-with-asthma.pdf

Bates, J., Goodale, P., Lin, Y., & Andrews, P. (2019). Assembling an infrastructure for historic climate data recovery: Data friction in practice. *Journal of Documentation*, *75*(4), 791–806. https://doi.org/10.1108/JD-08-2018-0130

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, *3*(2), 2053951716654502. https://doi.org/10.1177/2053951716654502

Borgman, C. L. (2015). What are data? In *Big data, little data, no data: Scholarship in the Networked World*. MIT Press. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7040577

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cancer Research UK, & Macmillan Cancer Support. (2016). *Improving awareness of the English cancer registry amongst patients, health professionals and the public*. https://www.macmillan.org.uk/documents/policy/improving-awareness-of-the-english-cancer-registry.pdf

Centre for Data Ethics and Innovation. (2020). *Review into bias in algorithmic decision-making*. Gov.UK. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making

Chico, V., Hunn, A., & Taylor, M. (2019). *Public views on sharing anonymised patient-level data where there is a mixed public and private benefit*. Health Research Authority. https://www.hra.nhs.uk/about-us/news-updates/sharing-anonymised-patient-level-data-where-there-mixed-public-and-private-benefit-new-report/

Connected Health Cities. (2017). *Connected Health Cities Citizens' Juries Report*. https://www.connectedhealthcities.org/chc-hub/public-engagement/citizens-juries-chc/citizens-juries/

De Roo, B., De Maeyer, P., & Bourgeois, J. (2016). Information flows as bases for archeology-specific geodata infrastructures: An exploratory study in flanders. *Journal of the Association for Information Science & Technology*, *67*(8), 1928–1942. https://doi.org/10.1002/asi.23511

Fabreau, G. E., Minty, E. P., Southern, D. A., Quan, H., & Ghali, W. A. (2018). A Metadata Manifesto: The Need for Global Health Metadata. *International Journal of Population Data Science*, *3*(1). https://doi.org/10.23889/ijpds.v3i1.436

Fiore-Gartland, B., & Neff, G. (2015). Communication, Mediation, and the Expectations of Data: Data Valences Across Health and Wellness Communities. *International Journal of Communication*, *9*(0), 19.

Ford, E., Kazempour, Y., Cooper, M. j. f, Katikireddi, S. v, & Boyd, A. (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*. https://doi.org/10.1136/bmjopen-2020-038006

Ford, E., Stockdale, J., Jackson, R., & Cassell, J. (2017). For the greater good? Patient and public attitudes to use of medical free text data in research: IJPDS (2017) Issue 1, Vol 1:229 Proceedings of the IPDLN Conference (August 2016). *International Journal of Population Data Science, 1*(1), Article 1. https://doi.org/10.23889/ijpds.v1i1.249

Gitelman, L., & Jackson, V. (2013). Introduction. In *" Raw Data " Is an Oxymoron* (pp. 1–14). The MIT Press.

Given, J., Nelson, E., Kane, F., Dolk, H., & Robinson, G. (2017). Public attitudes to data linkage and sharing. *International Journal of Population Data Science, 1:299 Proceedings of the IPDLN Conference (August 2016)*(1). https://doi.org/10.23889/ijpds.v1i1.320.

Hartman, T., Kennedy, H., Steedman, R., & Jones, R. (2020). Public perceptions of good data management: Findings from a UK-based survey. *Big Data & Society*, *7*(1), 2053951720935616. https://doi.org/10.1177/2053951720935616

Harwich, E., & Laycock, K. (2018). *Thinking on its own: AI in the NHS*. Reform. http://www.reform.uk/publication/thinking-on-its-own-ai-in-the-nhs/

Hays, R., & Daker-White, G. (2015). The care.data consensus? A qualitative analysis of opinions expressed on Twitter. *BMC Public Health*, *15*(1), 838. https://doi.org/10.1186/s12889-015-2180-9

Healthwatch England. (2018). *How do people feel about their data being shared by the NHS?* https://www.healthwatch.co.uk/report/2018-05-17/how-do-people-feel-about-their-data-being-shared-nhs

Kapofu, L. K. (2021). Researching the sociocultural: Modelling a responsive focused ethnography. *Methodological Innovations*, *14*(1), 2059799120987785. https://doi.org/10.1177/2059799120987785

Kennedy, H., Taylor, M., Oman, S., Bates, J., Medina-Perea, I., Ditchfield, H., & Pinney, L. (2021). *Living with Data survey report*. https://livingwithdata.org/project/wp-content/uploads/2021/10/living-with-data-2020-survey-full-report-final-v2.pdf

Kennedy, H., Taylor, M., Oman, S., Bates, J., & Steedman, R. (2021). *Public understanding and perceptions of data practices: A review of existing research*. https://livingwithdata.org/project/wp-content/uploads/2020/05/living-with-data-2020-review-of-existing-research.pdf

Kim, E., Rubinstein, S. M., Nead, K. T., Wojcieszynski, A. P., Gabriel, P. E., & Warner, J. L. (2019). The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology*, *29*(4), 354–361. https://doi.org/10.1016/j.semradonc.2019.05.010

Kitchin, R. (2014a). Conceptualising data. In *The data revolution: Big data, open data, data infrastructures and their consequences* (pp. 1–26). SAGE.

Kitchin, R. (2014b). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE.

Kitchin, R., & Lauriault, T. (2014). *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work* (SSRN Scholarly Paper ID 2474112). Social Science Research Network. https://papers.ssrn.com/abstract=2474112

MacIntyre, A. (2007). *After Virtue: A Study in Moral Theory, Third Edition*. University of Notre Dame Press. http://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=4454360

Martin-Sanchez, F. J., Aguiar-Pulido, V., Lopez-Campos, G. H., Peek, N., & Sacchi, L. (2017). Secondary Use and Analysis of Big Data Collected for Patient Care. *Yearbook of Medical Informatics*, *26*(1), 28–37. https://doi.org/10.15265/IY-2017-008

Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. U. (2017). Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. *Yearbook of Medical Informatics*, *26*(1), 38–52. https://doi.org/10.15265/IY-2017-007

Mouton, C., Baumann, H., & Biller-Andorno, N. (2018). Patient data and patient rights: Swiss healthcare stakeholders' ethical awareness regarding large patient data sets - a qualitative study. *BMC Medical Ethics*, *19*(1), 20. https://doi.org/10.1186/s12910-018-0261-x

Neves, A. L., Poovendran, D., Freise, L., Ghafur, S., Flott, K., Darzi, A., & Mayer, E. K. (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. https://doi.org/10.2196/14135

New Economics Foundation. (2010). *Who Sees What: Exploring public views on personal electronic health records*. New Economics Foundation. https://neweconomics.org/uploads/files/2cb17ab59382fe7c67_bfm6bdoas.pdf

Perera, G., Holbrook, A., Thabane, L., Foster, G., & Willison, D. J. (2011). Views on health information sharing and privacy from primary care practices using electronic medical records. *International Journal of Medical Informatics*, *80*(2), 94–101. https://doi.org/10.1016/j.ijmedinf.2010.11.005

Powell, A. (2019). Understanding and ethics. *Alison Powell*. https://www.alisonpowell.ca/?p=807

Schlegel, D. R., Ficheur, G., & Data", S. E. for the I. Y. S. S. "Secondary U. of P. (2017). Secondary Use of Patient Data: Review of the Literature Published in 2016. *Yearbook of Medical Informatics*, *26*(1), 68–71. https://doi.org/10.15265/IY-2017-032

Skovgaard, L. L., Wadmann, S., & 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*, *123*(6), 564. http://dx.doi.org/10.1016/j.healthpol.2019.03.012

Stockdale, J., Cassell, J., & Ford, E. (2018). *"Giving something back": A systematic review and ethical enquiry of public opinions on the use of patient data for research in the United Kingdom and the Republic of Ireland* (3:6). Wellcome Open Research. https://doi.org/10.12688/wellcomeopenres.13531.1

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

Tully, M. P., Bozentko, K., Clement, S., Hunn, A., Hassan, L., Norris, R., Oswald, M., & Peek, N. (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). https://doi.org/10.2196/jmir.7763

Understanding Patient Data. (2018). *Public attitudes to patient data use: A summary of existing research*. https://understandingpatientdata.org.uk/sites/default/files/2018-08/Public%20attitudes%20key%20themes_0.pdf

Vezyridis, P., & Timmons, S. (2017). Understanding the care.data conundrum: New information flows for economic growth. *Big Data & Society*, *4*(1). https://doi.org/10.1177/2053951716688490

Vikström, A., Moen, H., Moosavi, S. R., Salakoski, T., & Salanterä, S. (2019). Secondary use of electronic health records: Availability aspects in two Nordic countries. *Health*

Information Management: Journal of the Health Information Management Association of Australia, *48*(3), 144–151. https://doi.org/10.1177/1833358318817473

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, *20*(1), 144–151. https://doi.org/10.1136/amiajnl-2011-000681

Wellcome. (2015). *Wellcome Trust Monitor Report Wave 3 Chapter 6*. https://wellcome.org/sites/default/files/monitor-wave3-full-wellcome-apr16.pdf

Wellcome Trust. (2016). *The one-way mirror: Public attitudes to commercial access to health data*. https://wellcome.figshare.com/articles/journal_contribution/The_One-Way_Mirror_Public_attitudes_to_commercial_access_to_health_data/5616448/1