Participants' Attitudes Towards Data Sharing

Thesis submitted for the qualification of Doctor of Philosophy

Population Health Sciences Institute, Newcastle University

Nicola Louise Howe

December 2021

Abstract

Most research funders and journals now advise or require that research data should be made available for data sharing post publication. Sharing is widely practiced in other research communities but is less common in public health and clinical trials. When beginning this study, participants' views of data sharing in clinical trials or public health research were little explored.

This study therefore aimed to examine participants' attitudes towards data sharing where participants had taken part in clinical trials, public health research, longitudinal studies or were interested members of the public.

A questionnaire survey was developed, informed by a scoping focus group and a systematic review of the international literature. This was accompanied by a scoping review of grey literature.

Thematic analysis of the studies included in the systematic review identified six key themes and the grey literature review identified 16 relevant guidance documents. The questionnaire was completed by 1,664 participants. There was a large degree of corroboration between the systematic review data and the respondent's answers in the questionnaire. Generally, participants were most concerned about privacy and data security and exhibited concerns about open access and sharing with commercial organisations. Anonymisation and privacy were the areas of grey literature that converged the most with participant requirements, but generally the grey literature did not allude to participants' concerns. Recommendations for data sharing best practice were made, based upon the available evidence.

The strengths of this PhD study are the wide range of evidence gathered, however, as with many surveys, sample and response bias impact upon the generalisability of results.

This study provides up-to-date evidence from the UK regarding research participants' attitudes towards sharing of their study data. Researchers may use this data and the best practice recommendations to ensure that sharing practices align with participant preferences.

Acknowledgements

This study would not have been possible without the participants who took part. Participants from VOICE gave their time to assist with the scoping focus group and cognitive interviewing to help develop the questionnaire. Participants from VOICE, SAIL and SUPER participant groups, ACONF and ALSPAC took the time to complete my questionnaire and I am grateful for each and every response. The questionnaire would not have been as successful without the collaboration of ACONF and ALSPAC so a special thank you must go to those study teams for allowing my survey to be distributed to their participants.

This study could not have been completed, and would not even have begun, without my supervisors: Professor Elaine McColl, Professor Dorothy Newbury-Birch and Dr Thomas Chadwick, who have kept me going with help, feedback, and encouragement over the (long!) six or seven years that this study took. Thanks to Elaine in particular, for getting the PhD started back in the CTU days and for liaising with ALSPAC and getting the Data Access Agreement in place last year. Thank you also to Dr Emma Giles who was co-author on the systematic review and introduced me to coding in Nvivo. I am grateful to colleagues in Newcastle University's Population Health Sciences who have helped me over the years with systematic reviews, questions about PPI or have let me sit in to listen to taught modules.

Finally, a special thanks also goes to my family, friends and my husband who have listened to me go on about 'PhD stress' for a long time now and who eventually got to proof-read the thesis. Could not have done it without you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACONF	Aberdeen Children of the Nineteen Fifties
ALEC	ALSPAC Ethics and Law Committee
ALSPAC	Avon Longitudinal Study of Parents and Children
CASP	Critical Appraisal Skills Programme
CRD	Centre for Reviews and Dissemination
CRN	Clinical Research Network
CTU	Clinical Trials Unit
EPV	Events Per Variable
ESRC	Economic and Social Research Council
FAIR	Findability, Accessibility, Interoperability, and Reusability
FiCTION	Fillings in Children's Teeth, Indicated Or Not
GDPR	General Data Protection Regulation
HEFCE	Higher Education Funding Council for England
HRA	Health Research Authority
HTA	Human Tissue Authority
ICMJE	International Committee of Medical Journal Editors
ICO	Information Commissioners Office
IOM	Institute of Medicine
MRC	Medical Research Council
NCTU	Newcastle Clinical Trials Unit
NICA	National Innovation Centre for Ageing
NIHR	National Institute of Health Research
OCAP	Original Cohort Advisory Panel
ONS	Office of National Statistics
ORDTF	Open Research data Task Force
PIS	Participant Information Sheet
PPI	Patient and Public Involvement

QAS 99	Question Appraisal System
RQ	Research Question
SAIL	Secure Anonymised Information Linkage
SFG	Scoping Focus Group
SOP	Standard Operating Procedure
SUPER	Service Users for Primary and Emergency care Research
UKCRC	UK Clinical Research Collaboration
UKDS	UK Data Service
UKRI	UK Research and Innovation

Chapter 1 Background and Introduction

1.1 What is research data sharing?

'Data Sharing' refers to the process by which anonymised research data is made available at the end of a study, so that it may be utilised by other researchers, as recommended by funders, journals and research organisations (Chawinga and Zinn, 2019). Data may be shared so that further research can be conducted using the same data, data from separate studies can be combined for meta-analysis[1], the results of studies may be tested or replicated, and to provide complete transparency in the conduct of studies. Research conducted with data already collected is known as 'secondary research'. Data sharing occurs in many research disciplines such as the natural, social, medical, and genomic sciences (Tenopir *et al.*, 2011; Editorial, 2018). Within health research specifically, data sharing can be separated into sharing biological sample data (biobanks containing samples of blood or tissue) (Pereira *et al.*, 2014), routinely collected data (referred to as health records or GP data) (Casey *et al.*, 2016), clinical trial data (Lo, 2015) and health data linkage, (where separate datasets pertaining to the same individuals are linked) (Mourby *et al.*, 2019), with some overlap in practice between these categories.

This study will focus solely on sharing of health research data collected as part of a clinical trial, public health research or longitudinal study. Each of these types of research are briefly defined below.

Clinical trials are defined by the National Institute of Health Research (NIHR) as *"a research project that compares two or more treatments in patients with a particular condition or at risk of a condition to help generate high quality evidence about which is the more effective treatment or preventative strategy. The treatment being investigated in a clinical trial can be a medicinal product, a procedure, a device or another type of therapeutic intervention"* (NIHR, 2022a). Clinical trials provide the evidence base for identification of new or improved treatments, are highly regulated and are said to be the *"gold standard"* for evaluation of interventions (NIHR, 2022a).

[1] Deeks, J., Higgins, JPT., Altman, DG., (2021) 'Chapter 10: Analysing data and undertaking meta-analyses', in Higgins JPT, T.J., Chandler J, Cumpston M, Li T, Page MJ, Welch VA (ed.) *Cochrane Handbook for Systematic Reviews of Interventions version 6.2*.

Public health research is defined by the NIHR as *"practical interventions which have the potential to be delivered at scale, in order to generate evidence to support public health decision making and lead to sustainable population level change"* (NIHR, 2022b). Such research is carried out for example by Fuse (Fuse, 2022), the Centre for Translational Research in Public Health, which brings together North-East Universities in collaboration to *"deliver robust research to improve health and wellbeing and tackle inequalities"* (Fuse, 2022). Fuse work with the NHS, local and national government, voluntary and community sectors to help to transform public health via projects such as *"Energy drinks and young people's health"* and *"Drink Less app: reducing alcohol use"* (Fuse, 2022). While randomised controlled trials can be and are used to address research questions in public health, observational and quasi-experimental designs are often used in research in this area.

Longitudinal studies (often termed prospective cohort studies) are those where subjects are *"followed over time with continuous or repeated monitoring of risk factors or health outcomes, or both"* (Coggon *et al*., 2003) to determine any association between these risk factors and development of disease. Longitudinal studies vary in their size or complexity in terms of the number of follow ups received or information gathered but are *"generally observational in nature, with quantitative and/or qualitative data being collected on any combination of exposures and outcomes, without any external influenced being applied"* (Caruana *et al*., 2015, p. 2072).

These types of research were chosen to be the focus of this study for two reasons; the first is that these are the types of research with which I was familiar, they are conducted within the university setting in which I worked, and secondly, at the time of beginning this study there were far fewer research studies that focussed on sharing of these types of data. This is explored in more detail in section 1.7 below. It is recognised, nonetheless, that sharing of these types of data raises issues common to other categories of data sharing.

Data sharing is also sometimes referred to as 'open access' to data or 'open data'. The term 'open access' was first used to apply to journal articles which were available free of charge on the internet and is now also applied to study data (and accompanying documentation), which is available without charge and with no restrictions (Institute of Medicine (US), 2013; Attwood and Munafò, 2016). Data which are open access and have no restrictions placed on access or secondary use is often contrasted with data to which there is controlled access, i.e., where conditions are in place to control who can access it, and what secondary research

can be conducted (UK Data Service, 2021). Open data is said to increase transparency in research, both for researchers and the public (Institute of Medicine (US), 2013; Attwood and Munafò, 2016). Controlled access to data is said to ensure secondary research is scientifically sound, and that participants privacy is protected (Sydes *et al.*, 2015).

1.2 Data sharing requirements

As with other disciplines, sharing of health research data is recommended by stakeholders such as funders, journals, and research-supporting organisations (Walport and Brest, 2011; Ross and Krumholz, 2013; Taichman *et al.*, 2016).

For example, the Medical Research Council (MRC) make the following statement: *"The MRC expects valuable data arising from MRC-funded research to be made available to the scientific community with as few restrictions as possible so as to maximize the value of the data for research and for eventual patient and public benefit. Such data must be shared in a timely and responsible manner"* (Medical Research Council, 2016, p. 4).

Researchers must follow these stipulations because funders of research control access to research funding and journals control access to subsequent publication of study results. As exhibited in the quote above from the MRC, in health research increasing calls for sharing of study data have been coming from funders, journals and research organisations over the last ten to fifteen years, although Loder *et al* place the origins of the *"crusade"* for health research data sharing around 2003 with the National Institute of Health in the United States, followed by the journals Annals of Internal Medicine and the BMJ introducing data sharing statements in 2007 and 2009 respectively (Loder, 2013, p. 1). Borgman reminds us however, that UK funding agencies such as the Wellcome Trust and the Economic and Social Research Council (ESRC) began to formulate *"data release"* policies in the 1990s (Borgman, 2012, p. 1060). A more recent and obvious defining moment came from the *"joint statement of purpose"* published in the Lancet in 2011 (Walport and Brest, 2011, p. 537), where academic researchers, international organisations, funders, and funding agencies set out a vision to *"to increase the availability to the scientific community of the research data"* along with principles and immediate goals (Walport and Brest, 2011, p. 538). By 2013, the BMJ stipulated that all trials of drugs and medical devices would only be considered for publication if researchers committed to making the study data *"available on reasonable request"* (Godlee and Groves, 2012). This was followed by similar statements by the International Committee of Medical Journal Editors (ICMJE) who outlined *"the ICMJE's*

proposed requirements to help meet this obligation" (Taichman *et al.*, 2016, p. 467) along with separate statements from other journals encouraging placement of study data in repositories as a condition of publication (PLOS., 2014; Loder and Groves, 2015; Taichman *et al.*, 2016).

As a result of these requirements researchers need to incorporate plans and arrangements for data sharing into their trials and studies from the outset (Corti *et al.*, 2014), alongside robust data management and Good Clinical Practice (NIHR, 2016) by which high quality studies should be conducted. Wilkinson *et al* emphasise the importance of good data management and *"stewardship"* in order that data can be discovered and re-used in the future (Wilkinson *et al.*, 2016, p. 6). In brief, researchers need to pay attention to anonymisation, storage, access, and consent procedures. Ideally, before sharing, data should be anonymised (sometimes referred to as de-identified). Study data is usually pseudonymised; participants' names are replaced with a study number, but the data may contain potentially identifying information such as date of birth and ethnicity. Full anonymisation is a process by which the potentially identifying details such as date of birth, but also variables such as occupation, location and any free text fields are removed or amended to protect privacy (Institute of Medicine (IOM), 2015; Keerie *et al.*, 2018).

A suitable location (such as a repository) from which to share the data needs to be identified and researchers must decide whether data should be shared openly or whether some restrictions on access should be in place. Researchers also need to implement processes for secondary researchers to request data and set up a process or a panel to make decisions on sharing requests (Institute of Medicine (IOM), 2015; Cheah and Piasecki, 2020). Prior to this, participants should consent to their study data being shared, and ideally this should occur when consenting to the original research study (Corti *et al.*, 2014; Institute of Medicine (IOM), 2015; Ohmann *et al.*, 2017). Generally, participants are presented with one of two types of consent; broad one-off consent or re-consent (or what Corti *et al* term *"process consent"*) (Corti *et al.*, 2014). Broad one-off consent is given at the point of consent to the original research study whilst process consent is *"considered throughout the research project and assures active informed consent from participants"*- i.e.: consent when data is requested for sharing (Corti *et al.*, 2014, p. 25). The available guidance to which UK health researchers should adhere is explored in Chapter 3.

1.3 Evidence on health research data sharing

Broadly, there are two bodies of literature on the subject of health research data sharing: 'grey literature' such as policy documents and guidance on best practice for data sharing, and more traditional academic peer reviewed papers (which encompass opinion pieces, studies of research participants' attitudes, and recommendations for best practice). This study will utilise both types.

Throughout this PhD study I have scoped the grey and academic literature on health research data sharing, which informed both the grey literature scoping review (Chapter 3) and the systematic review (Chapter 2), a version of which was previously published as: (Howe *et al.*, 2018). Without attempting to repeat the content of chapters 2 and 3, a brief summary of that literature is presented below.

1.4 Rationale for data sharing

The advantages of data sharing are made clear, being extolled by funders, journals, and academics in the available literature e.g.: (Vickers, 2006; OECD, 2007; Walport and Brest, 2011; Ross and Krumholz, 2013; Editorial, 2018).

Often, the main benefit referred to is the chance for researchers to make use of data already collected to conduct their own separate research, increasing research efficiency, and providing greater benefits to science and ultimately leading to improved or more rapidly discovered treatments for patients (Mello *et al.*, 2013). The Wellcome Trust's view is of *"increasingly rich and complex datasets"* generated by the research they support as *"largely untapped resources of considerable value"* (Carr and Littler, 2015, p. 314). *"Novel approaches"* such as pooling or combining datasets can also result in *"the creation of datasets that permit a much wider range of research questions to be considered than can be addressed by the researchers who developed the methodology"* (UK Research and Innovation, 2018, p. 8). Existing studies can also be used as a basis for the design of new trials.

Data sharing is said to allow greater transparency in research, allowing *"close scrutiny of research results"* and verification of those findings, thereby increasing trust in the research process (Loder, 2013, p. 1). Transparency through shared data and associated documentation such as protocols can also lead to identification of weak trial design or operational errors, whilst new analyses can identify data errors or results previously un-

reported (Mello *et al.*, 2013). Building trust in the research process for participants is ever more important in a changing research landscape, for example, the introduction of the new GDPR (Information Commissioners Office, 2019b). Although GDPR does not apply to anonymous or anonymised data (which shared research data usually is), it has been observed that participants have an increasing awareness of their rights to control use of their data (Shah *et al.*, 2018), and that consent to access *"cannot simply be assumed based on failure to opt-out"* (Vlahou *et al.*, 2021).

Finally, data sharing ensures that researchers are meeting their ethical obligation to participants (Institute of Medicine (US), 2013; Mello *et al.*, 2013) with *"datasets collected at considerable expense using public and charitable research funds used in a manner that achieves the greatest possible benefit to health and society"* (Carr and Littler, 2015, p. 315). Not only are researchers getting the most out of participants' data, but by producing more research from shared data, participants will need to take part in fewer original studies, reducing the level of risk to which they are exposed (Mello *et al.*, 2013; Vallance *et al.*, 2016; Shabani and Obasa, 2019). Other work has identified that research paid for with public funds should also release their data for public scrutiny (Institute of Medicine (IOM), 2015; Attwood and Munafò, 2016) or *"public monies for public good rationale"* (Borgman, 2012, p. 1069).

1.5 Barriers to sharing

What Borgman in 2012 called the *"dirty little secret"* of data sharing is that it may not be happening very much at all in certain disciplines (Borgman, 2012, p. 1). Scientists are said to be habitually working much the same in the same way they have since the 18th century and to regard data as their own personal property (Boulton *et al.,* 2011) Whilst calling for research data to be *"made widely available to the research community"* Walport and Brest simultaneously identified that, although in fields such as genetics, molecular biology and social sciences, data sharing was well practiced, it had yet to *"be widely embraced by the public health research community"* (Walport and Brest, 2011, pp. 537-538). The same can be said for the clinical trial community (Hrynaszkiewicz and Altman, 2009). There seems to be a general consensus (Pisani *et al.*, 2010; Mauthner and Parry, 2013; Ross and Krumholz, 2013; Sturges *et al.*, 2015; Bouter, 2016) that there are barriers to be overcome or that data sharing is something to be strived towards rather than something that is happening regularly now. A later (2016) survey of clinical trials units (CTUs) themselves revealed that only 22% of those responding had a data sharing policy in place (Hopkins *et al.*, 2016).

Although writing directly in support of the ICMJE, in their 2014 systematic review of barriers to data sharing van Panhuis *et al* identified as many as twenty barriers to sharing (grouped into six main types; technical, motivational, economic, political, legal, and ethical) (van Panhuis *et al.*, 2014), three of which (ethical, technical, and professional) had been identified earlier by Pisani *et al* in their comment to the Lancet (Pisani *et al.*, 2010).

Other authors have focussed upon *"data management, data dissemination and validation of research contributions"* as challenges to sharing (Alter and Vardigan, 2015, p. 318). Data management or preparation of data (anonymisation) and documentation for sharing has resource implications (Carr and Littler, 2015; Hopkins *et al.*, 2016; MoreTrials, 2017; Devereaux, 2019). Metadata (explanatory data that accompanies the dataset) needs to be comprehensive enough that the secondary researcher does not need to contact the original researchers with questions (Alter and Vardigan, 2015). Good data management is therefore required from the outset of the study to reduce the burden on researchers at the point of sharing.

Alter and Vardigan identified dissemination as a participant issue, as participants can be anxious that their data could end up *"in the wrong hands"*, especially if data is shared without controlled access (Alter and Vardigan, 2015, p. 319) i.e., that secondary researchers may not respect the promises of confidentiality from the original researchers (Pisani *et al.*, 2010). From a researcher perspective, there may be a reluctance to disseminate datasets in which they *"invest significant time and effort in compiling…before they have had sufficient time to conduct and publish their own analyses"* (Carr and Littler, 2015, p. 314) whilst the *"direct costs (creating de-identified datasets, data dictionaries, data storage, and data security)"* and *"opportunity costs (…addressing questions related to trial datasets…to evaluate and process data requests)"* are said to divert money and researchers away from the business of actually conducting trials (Devereaux, 2019, p. 1).

Validation of research contribution refers to recognition of the original research team. Researchers are often reported to be concerned that they will not receive adequate recognition for collecting data subsequently used in secondary research (Pisani *et al.*, 2010; van Panhuis *et al.*, 2014) or that they will be *"scooped"* by a secondary researcher who will report findings before they are able to (Alter and Vardigan, 2015, p.320). Researchers may also find that they are not *"done"* with the data they are collecting (Borgman, 2012, p. 1069).

1.6 Facilitators of data sharing

Researchers (e.g. (Koers, 2016; Ohmann *et al.*, 2017; Hajduk *et al.*, 2019) have attempted to evaluate barriers and to suggest ways in which sharing can be facilitated or encouraged. Ohmann *et al* used a *"consensus building exercise"* with a *"stakeholder taskforce"* to produce 10 key principles (and 50 detailed recommendations) such as *"access to individual participant data and trial documents should be as open as possible and as closed as necessary…."* and *"the processing of data access requests should be explicit, reproducible, and transparent…"* (Ohmann *et al.*, 2017, p. 5). Other researchers have suggested that adjustments or incentives (such as funding instalments) should also be provided by funders and journals (Borgman, 2012; Bouter, 2016; Prisco *et al.*, 2016), for example, ensuring that (plans for) data sharing is a component of clinical trial registration and that main authors of trials are provided with an impact factor from studies using the data they collected (Prisco *et al.*, 2016, p. 1). Concerns could be addressed *"pre-emptively"* via the introduction of data sharing agreements (Polanin and Terzian, 2019).

Van Panhuis suggest that, to advance data sharing, all *"related barriers"* need to be addressed, rather than focussing on one single barrier at a time, and that *"specific data sharing strategies should be tailored to different types of data"* but it was beyond the scope of their review to suggest specifically how this could be achieved (van Panhuis *et al.*, 2014, p. 1475). Different approaches will similarly be required for different types of barrier.

Mauthner and Parry suggest that data sharing is a *"relational practice"* with relationships and trust between researchers more significant than improvements to infrastructure or policies, and that data sharing policies might do well to place more emphasis upon *"encouraging and facilitating"* rather than *"expecting and requiring"* (Mauthner and Parry, 2013, pp. 56, 62). Ohmann concur that *"no single group can be held responsible as the main drivers of data sharing"* and that each stakeholder must play their part (Ohmann *et al.*, 2017, p. 5). Hadjuk *et al* use a *"gap analysis"* to identify resources to encourage sharing given that a *"lack of clear standards and established guidelines"* in the first place is preventing sharing (Hajduk *et al.*, 2019, p. 1).

A key paper addressing data management concerns is that by Wilkinson *et al*, who in 2016, in response to the *"urgent need to improve the infrastructure supporting the reuse of scholarly data"*, published the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson *et al.*, 2016, p. 1). The FAIR principles were intended to enhance the

reusability of data held by researchers; data should be easy to find, accessible (for example, with protocol available), interoperable (for example using vocabularies that follow FAIR principles), and reusable (for example *"richly described"*) (Wilkinson *et al.*, 2016, p. 4).

Finally, Ross and Krumholz suggest that open science can be achieved by *"creating a culture that promotes sharing and provides credit to those who do—and consequences for those who do not"* (Ross and Krumholz, 2013, p. 1356).

1.7 Literature on participants' views of data sharing

Research which provides guidance on data sharing, such as Walport and Brest's (Walport and Brest, 2011, p. 537) *"high-level principles and goals"* to achieve greater data sharing, place emphasis on the research community rather than on patients or members of the public (i.e., the research participants) whose data are involved. In their paper on open access, Mauthner and Parry describe at length the resistance to sharing in academic communities and how this might be overcome through changes to infrastructure, ethics, and methodology (Mauthner and Parry, 2013) but typically, there is no mention of asking the participants how they feel about sharing.

There is now, however, a growing international body of literature on data sharing from the perspective of actual or potential research participants, but few of these papers are from the UK and few explore attitudes towards sharing of data from clinical trials, public health research or longitudinal studies specifically. Rather, much research explores public attitudes towards sharing of routinely collected health data (health records), biobank data and, to a lesser extent, linkage of various datasets (data linkage) (e.g. Stone *et al.*, 2005; Shabani *et al.*, 2014; Aitken *et al.*, 2016a; Graves *et al.*, 2019). In particular, there is a greater volume of literature concerning biobanks or sharing of biological samples, perhaps because sharing in the biological or genetic community started earlier than for other types of health data (Boulton *et al.,* 2011; Walport and Brest, 2011). A brief summary of existing research on participants' attitudes towards sharing of routinely collected health data, biobank data and data linkage is given below.

Studies exploring attitudes towards sharing of health data for secondary research have concluded that generally, participants are willing to share their data for altruistic reasons or to improve healthcare (Stone *et al*., 2005; Mazor *et al*.; 2017; Courbier *et al.*, 2019). However, participants have concerns about data security and the potential for exploitation

(Kass *et al.*, 2003; Mazor *et al*.; 2017) which leads them to desire a certain amount of control over use of their data (Courbier *et al.*, 2019) such as consent for secondary use sought in advance, at least as a courtesy (Nair *et al.*, 2004; Stone *et al*., 2005). Participants are reportedly more hesitant about sharing with commercial or profit-driven organisations or with organisations that they would not have chosen themselves than with universities for example (Stone *et al*., 2005; Kim *et al.*, 2015; Courbier *et al.*, 2019). Stone *et al* identified that participants were not clear what data may be shared, or why, or that sharing may already be occurring (Stone *et al*., 2005).

Similar to participants questioned about sharing of health data, participants who have donated blood or tissue are reported to be happy to share their data with biobanks for secondary research (Ludman *et al*., 2010; Treweek *et al.*, 2009; Shabani *et al.*, 2014; Joly *et al*., 2015) but also have concerns about privacy and confidentiality (Kaufman *et al.*, 2009; Lemke *et al*., 2010; Chan *et al.*, 2012; Shabani *et al.*, 2014) and that their information could be used against them (Kaufman *et al.*, 2009; Chan *et al.*, 2012; Shabani *et al.*, 2014), for example through *"genetic discrimination"* (Lemke *et al.,* 2010, p. 369). Trust in the organisation performing the research is also key, with pharmaceutical or for-profit companies identified as less trustworthy (Lemke *et al*., 2010; Shabani *et al.*, 2014). A US based systematic review of attitudes towards consent for de-identified human data and specimens to be included in a biobank found that *"willingness for data to be shared was high, but it was lower among individuals from under-represented minorities, individuals with privacy and confidentiality concerns, and when pharmaceutical companies had access to data"* (Garrison *et al*., 2016, p. 663).

Fewer studies examining attitudes towards data linkage were identified during the course of this study. Clarke *et al.* found that more than half of their participants would be willing to link their health records to lifestyle data, but participants did have concerns about data getting into the *"wrong hands"* i.e., someone other than the secondary researchers and the potential for invasion of privacy (Clarke *et al.*, 2022, p. 13). Xafis found that contrary to their assumption, most participants in most scenarios would not require to give consent for linkage as long as data was de-identified as once data had lost its identifiers it became *"completely detached"* from individuals (Xafis, 2015, p. 6). However, when a scenario whereby researchers and not 'experts' linked health, work and employment data, participants preferred to be asked for consent (Xafis, 2015). In interviews with participants

aged 17-19 Audrey *et al.* identified that although attitudes towards health research were positive, participants were more comfortable with use of some data (e.g., asthma, heart disease) than other data (e.g., teenage pregnancy, mental health) which might be stigmatising (Audrey *et al*., 2016). Consent was seen to be synonymous with individual 'opt in' consent and even if data were anonymised this was not seen to negate the requirement for consent for linkage, as consent should be offered both as a courtesy, and so that the participant could choose to prevent what they saw as the risk of (inadvertent) disclosure (Audrey *et al*., 2016). Approximately 46% respondents to Aitken *et al*'s questionnaire decided that data linkage for research purposes was *"unacceptable under any circumstances"* (Aitken *et al*., 2018, p. 11). Overall, the factors identified as most likely to influence Aitken's respondents' preferences were the type of data being linked and 'how profits are managed and shared' (Aitken *et al*., 2018). Differences in attitude were found to be linked to age, gender, health, and employment.

A systematic review plus focus group study published in 2013 regarding participants' understanding of sharing of primary or secondary care health records produced 27 relevant articles, six of which were from the UK (Hill *et al.*, 2013). Hill *et al* concluded that participants who were older or male were more likely to consent to review of health records for research, with all participants keen to contribute to research but cautious about data misuse and commercial gain from their data (Hill *et al.*, 2013). More recent reviews of attitudes towards sharing or linkage of health data (Aitken *et al.*, 2016a) and sharing of 'health data' (Kalkman *et al.*, 2019a) also found 25 and 27 papers respectively, some of which appeared in both or in the earlier review from Hill *et al*, demonstrating that there is a lot of overlap between research into sharing of biobank data, heath records and data linkage. Often, systematic reviews into data sharing encompass papers from all areas of sharing, presumably as there are not enough papers in each distinct area to inform a systematic review, or because researchers assume that participants' attitudes will be similar regardless of the type of sharing. Aitken *et al's* review (Aitken *et al.*, 2016a) included papers exploring attitudes towards linkage of data sets as well as sharing of health records and genomic data. Aitken *et al* identified seven themes which can be summarised as participants exhibiting a *"general-though conditional-support for linkage and sharing for research purposes"* and concerns regarding misuse of data, confidentiality, and control (Aitken *et al.*, 2016a, p. 1). Kalkman *et al's* review (Kalkman *et al.*, 2019a) included papers on sharing of data from

clinical studies, health record data, genetic data, data linkage and studies exploring attitudes towards consent for sharing. Results supported those of *Aitken et al* in that participants' saw benefits of sharing but expressed concerns about *"breaches of confidentiality and potential abuses of the data"* (Kalkman *et al.*, 2019a, p. 1). A further review which incorporated papers on *"health consumers"* views of sharing health record and clinical trial data and excluded views on biobanking and genetic research and health records where possible, identified a total of 75 papers, 35 of which focussed on consumers' concerns regarding *"privacy, trust and transparency"* (Hutchings *et al.*, 2020, p. 1). When the 35 studies were synthesised, the authors concluded that participants wanted a balance between public benefit and individual privacy (Hutchings *et al.*, 2020).

1.7.1 Consent

Many of the studies exploring attitudes towards sharing of routinely collected health data, biobank data and data linkage identified that participants wanted to consent to sharing (Nair *et al.*, 2004; Stone *et al*., 2005; Xafis, 2015; Audrey *et al*., 2016; Garrison *et al*., 2016; Kalkman *et al.*, 2019a). This begs the question as to whether we consider that the consent given at the beginning of the original study is sufficient. To ensure that consent is informed, it must be freely given with sufficient information provided on all aspects of participation and regarding present or future data use. Participants may be asked to consent for use of their data in future studies as well as the one in which they are currently consenting to, but do they realise this, will they remember this, and should consent be sought again in the event of future research? Prisco *et al* suggested researchers could use an *"extended informed consent"* where participants are free to agree or decline use of their data for further clinical studies (Prisco *et al.*, 2016, p. 1). Some research (albeit concerning biobanks) has identified that participants would prefer to be contacted to re-consent before their data were used in further research (Robling *et al.*, 2004; Lemke *et al.*, 2010; Ludman *et al.*, 2010), and that this re-consent might be conditional, depending on the type of organisation with which data is to be shared (Trinidad *et al.*, 2010; King *et al.*, 2012; Grande *et al.*, 2013; Hill *et al.*, 2013). Nontheless, some research has reported that participants are happy to give consent for sharing once, with the consent for the original study (Campbell *et al*., 2007; Clerkin *et al*., 2013; Braun *et al*., 2014), or are willing to accept a broad consent model even if this is not their first preference (Taylor and Taylor, 2014).

1.7.2 Access types

In their systematic review of patient attitudes towards sharing of biological data, Shabani *et al* identified the concept of control as being important to participants: *"people have a right to control their information. It doesn't matter whether anything bad would happen"* (Shabani *et al.*, 2014, p. 6). An oft-visited aspect of control is type of access to be used for data for secondary use. Shabani and Obasa explore types of access and associated policy used by industry-sponsored trials (Shabani and Obasa, 2019), whilst Sydes *et al* (Sydes *et al.*, 2015) and Tucker *et al* (Tucker *et al.*, 2016) advocated for controlled access with their guiding principles for controlled access and recommendations for best practice respectively. In the systematic review (Chapter 2) participants' views on access types are explored, and in Chapter 3 published guidance on access is summarised.

1.8 The aim of this research study

As described above, although health research data sharing is a growing area in research and in guidance and indeed has increased in prominence over the course of this PhD study, there is still a lack of distinct research on the views of participants from clinical trials, public health research or longitudinal studies towards data sharing (Mello *et al.*, 2018). There is often no clear distinction between attitudes towards sharing of these types of data and attitudes towards record linkage, sharing of routinely collected (health) data or biobank data. Instead, all these types of sharing are often combined or conflated as occurred in the papers described above by Hill *et al*, Aitken *et al* and Kalkman *et al* (Hill et al., 2013, Aitken et al., 2016a, Kalkman et al., 2019a). There is also a more general lack of research originating in the UK.

Although existing research is available into participant attitudes towards sharing of biological data or health records for research, some of which has been summarised or referred to above, it is possible that attitudes of participants taking part in clinical trials, public health research or longitudinal studies will be different. Indeed, the specific types of question that need to be asked of these participants will also be different due to the operational nature of clinical trials, public health or longitudinal studies as compared, for example, to allowing access to health records for secondary research.

The overall focus of this study was therefore research participants' attitudes towards health research data sharing, particularly in respect of data from clinical trials, public health

research and longitudinal studies. From this point on, and in remaining chapters, this kind of health research data sharing will be referred to as 'data sharing' or 'sharing' for brevity. When talking about other types of sharing in the literature, for example sharing of biological data, this distinction will be made clear.

The specific research questions of the PhD study are:

1. What are participants' attitudes towards data sharing (and how may these differ according to socio-demographic characteristics and prior research experience)?
2. Does knowing that their data may be shared affect their likelihood to participate in research?
3. What are their preferences regarding data sharing?
4. To what extent does current guidance reflect research participants' views and priorities?

The methods employed to answer these research questions are:

- A systematic review of existing international literature on research participants' views of health research data sharing;
- A scoping review of available grey literature, in the form of UK guidance documents on health research data sharing; and
- A questionnaire survey to measure attitudes of UK research participants and members of the public towards health research data sharing.

After gathering evidence on attitudes towards sharing from the systematic review and the questionnaire survey and then comparing this to current best practice guidelines I will make my own recommendations for best practice when sharing data from participants of clinical trials, public health, or longitudinal studies.

1.9 Epistemological position

Prior to beginning this research study, I needed to consider my epistemological position; that is, what theoretical approach I would take to research, and the beliefs that underpin this approach. Green and Thorogood describe epistemology as the *"theory of knowledge"* that shapes how we see the world and understand health knowledge (Green & Thorogood, 2014, p. 11). Accordingly, the researcher's epistemological position affects not only how the results of research are interpreted, but how the research is conducted in the first place.

Broadly, traditional medical 'science' and research, resides within the realm of positivism, whereby the world and phenomena can be measured or studied until the truth is identified or *"there is a stable and knowable reality, separate from our human understandings of that reality"* (Green & Thorogood, 2014, p. 13) but that we can get there with investigation. By contrast, qualitative research can often be viewed as taking a more interpretive approach, that there is no one 'correct' answer, and the right answer depends on who or how you ask; in other words, it seeks to *"understand human behaviour"* rather than *"explaining people or society"* (Green & Thorogood, 2014, p. 13). Interpretivist approaches can be described as *"a response to the over dominance of positivism"* in science (Grix, 2010, p. 83) and are also aligned with social constructionism wherein reality is 'socially constructed', a result of processes and understandings, and therefore will differ between individuals or societies (Green & Thorogood, 2014, p. 17). Grix counters the either/or approach to research described above by explaining that there is also a third research paradigm, *"post-positivist"* which falls somewhere in between interpretivism and positivism, a *"critical realism"* that asks both *"how?"* and *"why?"* (Grix, 2010, pp. 79-84). Both positivism (and post-positivism) and interpretivism are umbrella terms for a whole range of approaches to and beliefs about research.

As explained above, the type of research conducted, and the research paradigm are intertwined. My research uses a mix of both qualitative and quantitative methods. Mixed method research, or *"triangulation"* (Grix, 2010, p. 136), defined as researchers collecting and analysing both quantitative and qualitative data within the same study, involves a purposeful *"mixing of methods in data collection, data analysis and interpretation"* (Shorten and Smith, 2017, p. 74) to answer the same question. Triangulation is *"often used to describe research where two or more methods are used, known as mixed methods"* and the use of both qualitative and quantitative methods can either lead to the same conclusions being drawn from both methods or the results from both methods being *"complimentary"* or divergent (Heale and Forbes, 2013, p. 98).

Mixed methods research can be grounded in either *"a-paradigmatic stance, the multiple paradigm stance or the single paradigm stance"* as, given the mix of both qualitative and quantitative approaches, mixed-methods do not fit neatly into a single paradigm (Hall, 2013, p. 2). Rather than advocate for a single paradigm to *"legitimise"* the approach, Hall argues for a grounding in a *realist* perspective (Hall, 2013, p. 2).

However, although this study used both qualitative and quantitative methods of investigation, I had not at first intended to be a 'mixed methods' piece of research, perhaps because at that time the definition of mixed methods to me was not clear (Johnson *et al.*, 2007). The qualitative aspects of the research (scoping focus group, inclusion of qualitative literature in the systematic review) served to inform the quantitative questionnaire survey; at first, I therefore viewed the qualitative aspects as a means to an end. Perhaps I intended *"multimethod"* research (Shorten and Smith, 2017) as I was choosing the most appropriate method for each component question within the same epistemological approach. Mixed methods could actually provide me with *"the most informative, complete, balanced, and useful research results"* (Johnson *et al.*, 2007), and combining qualitative and quantitative research methods together in a sequence as I have done, is referred to as 'sequential mixed methods' (Tashakkori and Teddlie, 2010; Edmonds and Kennedy, 2017; Creswell and Plano Clark, 2018). I have therefore settled on a sequential mixed methods approach for this study.

Sequential mixed methods is an approach by which researchers deploy qualitative and quantitative methods in sequence, for example using qualitative methods to collect some data, analysing the results of that phase, and then using those results to direct or inform the next quantitative phase of research, for example focus groups informing a questionnaire survey as I have done here. The emphasis may still be upon either the quantitative or the qualitative aspect of the research project (Tashakkori and Teddlie, 2010; Edmonds and Kennedy, 2017) and either the qualitative or quantitative research may come first (Ivankova *et al.*, 2006). The way in which the two components of the research are combined must also be decided by the researcher (Ivankova *et al.*, 2006). The rationale for this sequential approach lies in *"first exploring a topic before deciding what variables need to be measured"* (Edmonds and Kennedy, 2017, p. 3). Edmonds and Kennedy talk about using the results of a literature review alongside qualitative research, and how the qualitative research does not make the results of the literature review any less valid, but simply provides additional evidence. I am conducting sequential qualitative-quantitative analysis.

Given that the primary focus of this study was to be the questionnaire survey it might be expected that my epistemology was positivist (given the quantitative nature of the questionnaire). However, I also used qualitative methods (scoping focus group and qualitatively presented narrative synthesis in the systematic review and scoping review). I felt that I was approaching this study with a focus on interpretation, though not discounting

that there are unobservable processes at work behind these interpretations of participants' observed preferences. The systematic review gathered the evidence but did not provide an explanation for findings. It was a narrative synthesis, not a more 'measurable' (and therefore positivist) meta-analysis. The grey literature review was scoping and exploratory in nature, with the caveat that the presentation of results was my interpretation, within the framework of topics identified in the systematic review, of the available literature meeting the inclusion criteria. Although the questionnaire survey used quantitative methods for collection and analysis of data, and the results are presented quantitatively, the data collected were the subjective views of participants. It was really the respondents' opinions and the reasons for these that I was interested in, in other words, how the participants' experience of the world and of research affects their views on data sharing. The interpretivist, subjective explanation for participants' attitudes towards data sharing may well result in a conclusion that is *"open ended rather than complete"* (Grix, 2010, p. 83). The set of recommendations I give at the end will be my interpretation of participants preferences. Despite this, I still intend that all methods should be clearly set out and reproducible.

Tashakkori and Teddlie, although arguing that research methods should not be firmly tied to any one epistemological stance, have summarised that the position of a researcher using sequential concurrent mixed methods is likely that of a 'critical realist' or simply, 'realist' (Tashakkori and Teddlie, 2010). According to critical realism, there is *"a reality that exists independent of our thoughts about it, and while observing may make us more confident about what exists, existence itself is not dependent on observation"* (Haigh *et al.*, 2019). This is a mid-point between interpretivism and positivism. There is a reality to measure, but what is observed or measured can change over time as our understanding changes. For mixed methods research, realism is argued to *"validate and support key aspects of both qualitative and quantitative approaches while identifying some specific limitations of each"* (Tashakkori and Teddlie, 2010, p. 2). This fits with my interpretivist stance; I am happy to position myself within a realist stance. For me, this means that my results will be subjective, based upon the participants who are reporting their opinions, and the influence or bias that I bring in summarising them, but that there are reasons behind the observed preferences and opinions that may not be fully measured or observed as part of this research. Further details on critical realism can be found from Tashakkori and Teddlie who provide an interesting

exploration of realism in the context of mixed methods research (Tashakkori and Teddlie, 2010).

1.10 My position in research

It is also important to briefly consider my position in the world of research whilst conducting this study, as this may also influence my methods and interpretation of findings. For the duration of the PhD study, I have worked as a database manager at Newcastle Clinical Trials Unit, part of Newcastle University. This involves dealing with participants' data on a daily basis. Although this data is pseudonymised and I do not have any contact with the participants themselves, I am conscious when using this data that it represents a real person's contribution to a study, and that I am respecting their privacy by following GDPR and data protection regulations as a matter of course. In recent years the unit has begun to receive an increasing number of data sharing requests, and preparations are being made for long term accommodations for sharing, for example consent forms now include a sentence about sharing, rarely included in older studies than began five or ten years ago. Data are being prepared for sharing through anonymisation, both as a standard for new studies and on an ad-hoc basis for older studies without the specific resource in place to do this. The unit's response to sharing, as per the literature is often secondary to the everyday trials work of the unit. In addition to this role, since 2015 I have been a member of the data sharing task and finish group set up by UK Clinical Research Collaboration (UKCRC) Registered Clinical Trials Units, and we have recently published a guidance Standard Operating Procedure (SOP)[2] that may be utilised by units who do not yet have one. I am therefore aware of some of the challenges posed by sharing from a researcher perspective but chose to advocate for participants with my PhD research. I am hoping that some of the recommendations made here, on behalf of participants, may even be applied to processes and practice within the trials unit where I work.

1.11 Research technique rationale

This PhD study utilised a questionnaire survey to address the main research question – what are participants' attitudes towards data sharing? But the survey required development and testing prior to its distribution to participants. Recognising that the focus of the survey was the views of research participants and members of the public regarding data sharing, and

[2] https://ukcrc-ctu.org.uk/wp-content/uploads/2021/04/data_sharing_sop_guide_v1.1_-1.pdf

that such individuals were the target 'audience' for the questionnaire, it was considered essential to include the perspective of such individuals in establishing questionnaire content. This accords with principles for good questionnaire design. When describing a series of 3 interviews to develop questionnaires, Stettler and Featherstone neatly summarise the process in 3 key principles; *"Start fresh"*, *"Learn the respondent's language"* and *"Know thy user, for it is not you"* (Stettler and Featherston, 2012, p. 1). This translates to starting with no predetermined questionnaire, using language suitable for the expected respondents, and ensuring that the questionnaire is usable and easy to navigate.

For the current study, the participant perspective and orientation in questionnaire content was achieved in two ways. First, through the interrogation of the existing international literature to identify areas of concern for research participants and members of the public regarding research data sharing (Chapter 2) and thereby inform questionnaire content. Second, through a scoping focus group which elicited the attitudes of a group of research-interested individuals in the North East of England. Both the scoping group work and subsequent cognitive interviews (see Chapter 4) ensured that the questionnaire survey was not just informed by current literature but was reflective of areas of concern and interest of research participants and members of the public, whilst being understandable and easy to navigate.

Accordingly, this involvement of (interested) members of the public in the development of the questionnaire went further than cursory patient and public involvement (PPI); I considered it to be more representative of 'co-production' techniques (Newbury-Birch and Allen, 2019; UK Research and Innovation, 2021). Co-production is described as *"involving people from outside the research community"* in research, either at design stage or in delivering the research project (UK Research and Innovation, 2021) and as *"exchange, synthesis, and dissemination of knowledge between researchers, policymakers, and end users"* (Newbury-Birch and Allen, 2019, p. 2). Public and/or participant involvement in designing research can confirm that the study *"best addresses the needs of individuals and communities"* (UK Research and Innovation, 2021). It was anticipated that the questionnaire survey, combined with the systematic review, would reveal enough about participants' attitudes and preferences that my own recommendations for best practice could be made at the end of the thesis. Newbury-Birch and Allen explain that one of the key reasons for undertaking co-production research is to be able to influence policy and practice (Newbury-

Birch and Allen, 2019, p. 10), and it is with the recommendations for best practice that I hope to be able to do this. Participants may also gain something from being invited to take part in research in this way. Being invited as experts and collaborating with researchers can be *"empowering"* (Gibbs, 1997, p. 3). The ways in which participants co-produced the content and layout of the questionnaire is described in Chapter 4.

The structure of the remainder of this thesis and how each research question is answered is outlined below:

1.12 Outline of thesis:

This chapter has attempted to situate this PhD study in the context of the current research landscape by providing a brief overview of the types of literature available, the broad topic areas this literature covers and the research gap that remains; namely the attitudes of UK participants towards health research data sharing, where the participants have taken part in a clinical trial, public health or longitudinal study or are potential participants (members of the public) in such a study. I also outline my epistemological position, the rationale for my choice of research methods and how my position in research influenced my choice of study.

Chapter 2 contains the systematic review; an exploration of existing international literature regarding participants' attitudes towards health research data sharing, where participants were taking part of an ongoing study or were members of the public as potential participants. Studies focusing on clinical trials or public health research were prioritised. Both qualitative studies (e.g., using focus groups or interviews to elicit knowledge and attitudes) and quantitative studies (e.g., using questionnaires) were accepted. The systematic review explored research questions 1-3 as far as possible, presenting themes identified during a thematic synthesis of the literature.

Chapter 3 contains the grey literature review which identifies and collates all relevant UK guidance, regulations, or best practice documents for researchers regarding data sharing. This chapter addresses research question 4. These guidance documents were searched for specific guidance on the same areas of concern or interest identified in one or more of the themes of the systematic review.

Chapter 4 details the development of the questionnaire survey, through co-production techniques such as use of a scoping focus group, readability testing and cognitive interviewing, through to questionnaire distribution.

Chapter 5 picks up where Chapter 4 leaves off, with a completed questionnaire draft ready to distribute. This chapter explains how the questionnaire was distributed including the sampling strategy, data cleaning and manipulation once completed questionnaires were received and how analysis was conducted. This chapter concludes with presentation of the summary results of the questionnaire survey and significant results from secondary analyses. This analysis attempted to answer questions 1-3.

Chapter 6 concludes the thesis with a summary of all results obtained from the systematic review, grey literature review and questionnaire survey and a comparison of participants' attitudes compared to best practice guidance. This comparison is structured to echo the layout of the questionnaire with distinct topic areas addressed one by one. This is followed by an acknowledgement of the strengths and limitations of the PhD study, my own recommendations for best practice in future research and finally, any areas of research that are outstanding, based upon the results of the questionnaire survey, and supported by the systematic review findings.

Chapter 2 Systematic Review of Participants' Attitudes Towards Data Sharing: A Thematic Synthesis

2.1 Introduction

This chapter details the methods and findings of a systematic review of the international literature on participants' attitudes towards health research data sharing. An earlier and more concise version of this chapter, including the 9 papers published and screened prior to March 2018, has been published as a peer reviewed paper (Howe *et al.*, 2018) in the Journal of Health Services Research and is available at: https://journals.sagepub.com/doi/full/10.1177/1355819617751555 and in Appendix A.

2.2 Background

In light of increasing requirements by funders and journals to share research study data, a resultant shift to a sharing culture amongst researchers (Chapter 1 Background and Introduction), and guidance detailing how they might better share data (see Chapter 3 Grey Literature Review), there has been a subsequent increase in academic literature exploring participants' views of data sharing, although this is primarily focussed upon sharing of biobank data or health record data (Chan *et al.*, 2012; Hill *et al.*, 2013). There is a limited amount of literature on the perspectives on data sharing of those participating in clinical trials, public health, or longitudinal studies.

To address the first objective of the current doctoral study (Chapter 1 Background and Introduction section 1.8), it was necessary to identify and synthesise the available research on participants' attitudes towards research data sharing, focussing specifically on data from clinical trials, public health research or longitudinal studies. By collating and synthesising all available evidence on this topic, it becomes more accessible and usable for researchers. In addition to this, collating all existing research also increases the reliability of any conclusions drawn from the evidence and helps to identify any evidence gaps (Centre for Reviews and Dissemination (CRD), 2013). Systematic reviews provide a systematic way of collating this evidence using *"scientific…explicit, pre-specified and reproducible methods"* (Centre for Reviews and Dissemination (CRD), 2013, p. V).

There are varying approaches to analysing data captured during the systematic review process, depending on the type of evidence (data) gathered. Petticrew and Roberts (Petticrew and Roberts, 2006) broadly split these into meta-analysis for quantitative data

and narrative or meta synthesis for qualitative or social sciences data. There are, however, many variously named, interlinked and interchangeable methods for analysis of qualitative data, only one of which is referred to as narrative synthesis (for example as described by Thomas and Harden in 2008 (Thomas and Harden, 2008)). An overview of these qualitative synthesis methods and their epistemological position is provided by Hannes and Lockwood (Hannes and Lockwood, 2011). The method selected for this systematic review is detailed in the section below on data synthesis and analysis.

The aim of this systematic review was to examine, using systematic reviewing methods, the international literature on research participants' attitudes towards data sharing in the context of clinical trials and other public health research. It specifically explores participants' attitudes towards sharing, and whether awareness of data sharing might affect consent to take part in research.

2.3 Methods

A protocol for the review was developed using the PRISMA-P (Preferred reporting items for systematic review and meta-analysis protocols) 2015 checklist (Moher D, 2015) and followed throughout the systematic review process. The protocol was not eligible for PROSPERO registration as the systematic review did not focus on health outcomes (PROSPERO, 2017). During final review of this chapter, the updated PRISMA 2020 checklist was studied to ensure that there were no new elements of the checklist that had not already been included in this review (Page *et al.*, 2021). No additional outstanding information was identified, and the decision was made to keep the original PRISMA flow diagram which had been adapted already to incorporate an update to the review. There were no protocol deviations when conducting the review.

2.3.1 Search strategy

I piloted search terms in a Medline scoping search which returned few relevant studies with the use of either broad or narrow search criteria. When conducting this review, the search terms were therefore left broad, to maximise the number of studies included in this relatively under explored area (see terms in Appendix B). Terms relating to data sharing and participant, patient or public attitudes were used to interrogate the following databases: Medline, Embase, Web of Science, ASSIA, CINAHL, HMIC and PsychINFO. Key search terms were taken from studies already identified serendipitously or through personal contacts and

were adapted for each database. The database searches were refined until they picked up those relevant papers already identified serendipitously. Letters to the editor, books, conference proceedings and editorials were excluded. The search was restricted to studies concerning 'humans'. Reference and citation lists of included studies, publications of included first authors and references within systematic reviews were also searched. For several reasons, existing systematic reviews were excluded from the inclusion criteria set out in the protocol. Firstly, it was anticipated that there might be some difficulty in extracting sufficient data regarding participants' attitudes from a review which had already summarised and condensed a number of studies. By including systematic reviews, I would also be required to interpret and code data which had already been interpreted and re-presented by the author of the review, thereby increasing the opportunity for misinterpretation. In addition, studies included in any systematic review may have had differing methodologies or subject matter to those included in this review and may therefore provide data which would not otherwise be considered. Instead, the individual studies included in any identified systematic reviews were assessed for suitability for inclusion if they had not already been identified as part of the systematic database search.

Post publication of the original systematic review (Howe *et al.*, 2018), and over the course of the remainder of this PhD study, I continued to receive periodic journal database alerts with the up-to-date results of my saved searches (from Medline, Embase, Web of Science, ASSIA, CINAHL, HMIC and PsychINFO). Each email alert therefore contained potentially eligible papers. The review was updated to include papers from alerts up to and including 5th August 2020.

2.3.2 Inclusion and exclusion criteria

To be included, studies had to report qualitative, quantitative or mixed-methods empirical research. They had to address data sharing, more specifically regarding secondary use of research data already collected as part of a trial, study, or intervention. Included studies also had to examine attitudes of research participants or potential participants, i.e., members of the public. They had to be published after 1995 (year of publication of EU Directive 95/46/EC; the Data Protection Directive) (Data Protection Commissioner, 1995).

Studies concerning sharing of biobank data, human tissue, blood samples, routinely collected primary and secondary care data (health records) or 'data-linkage' were excluded. The systematic review protocol had originally allowed inclusion of these types of studies

should the identified number of studies on sharing of clinical trial or health intervention or longitudinal study data have been too low. During screening it was decided that this was not the case. There were no restrictions on language or country of origin.

2.3.3 Study selection

All potentially eligible titles identified during database searches (n=16,318) were downloaded as citations into EndNote (EndNote, 2021). Additional records (n=787) identified through reference and citation lists of included studies, publications of included first authors and references from systematic reviews were also downloaded as citations into EndNote. All records were then de-duplicated.

For preparation of the initial, published systematic review, I screened all titles and one of my supervisors (DNB) acted as second reviewer, independently screening 20% of all titles. Although there was a great deal of consensus, we erred towards inclusion if uncertain whether an article was eligible. Most titles were unsuitable and were rejected (n=16,176) because they did not relate to data sharing at all and had been included in database searches because the titles or abstracts contained phrases such as 'participant attitudes', 'privacy' or 'health data' for example. Both I and my supervisor (DNB) then independently screened abstracts for all accepted titles (n=161) against the inclusion criteria, noting our reasons for exclusion. At this stage, most papers were excluded because they explored attitudes towards biobanking in general (n=23), sharing of biological data (n=22), sharing of health records (n=24) or did not contain empirical research (n=25) (see Figure 1). Again, if eligibility was uncertain, for example if an abstract was not immediately available or it was unclear from the abstract whether the study met my inclusion criteria, it was retained and included. Reconciliation of disagreement was achieved through discussion and by erring towards inclusion. Rejected titles and papers from all stages were saved in separate EndNote files. Papers that remained after abstract screening (n=51) were read in detail by myself, with later group discussion including two of my supervisors (DNB & EM).

Screening of papers in the journal email update alerts from 2018 to 2020 was completed by me, as and when the emails arrived. Titles were screened, and any titles that potentially met the inclusion criteria were downloaded and saved for abstract screening (n=28). If it was not clear after abstract screening, the full paper was then accessed and read, to see if it met the inclusion criteria. Rejected papers were saved with reasons for rejection. The references, citation lists, and first-author publications of papers identified through email alerts were not

formally checked for additional potential papers. This is because, post publication of the original systematic review (2018 onwards), the amount of literature on data sharing had increased, was easier to find, and was observed to be more likely to refer specifically to 'data sharing' in its keywords or title (meaning it would be more likely to appear in my email alerts). If any of the authors of newly identified eligible papers referred in the text to a potentially eligible or interesting paper of which I was not already aware, this was sought for screening. The PRISMA (Moher *et al.*, 2015) diagram in Figure 1 details this screening process for both the original database searches and for the subsequent email alerts.

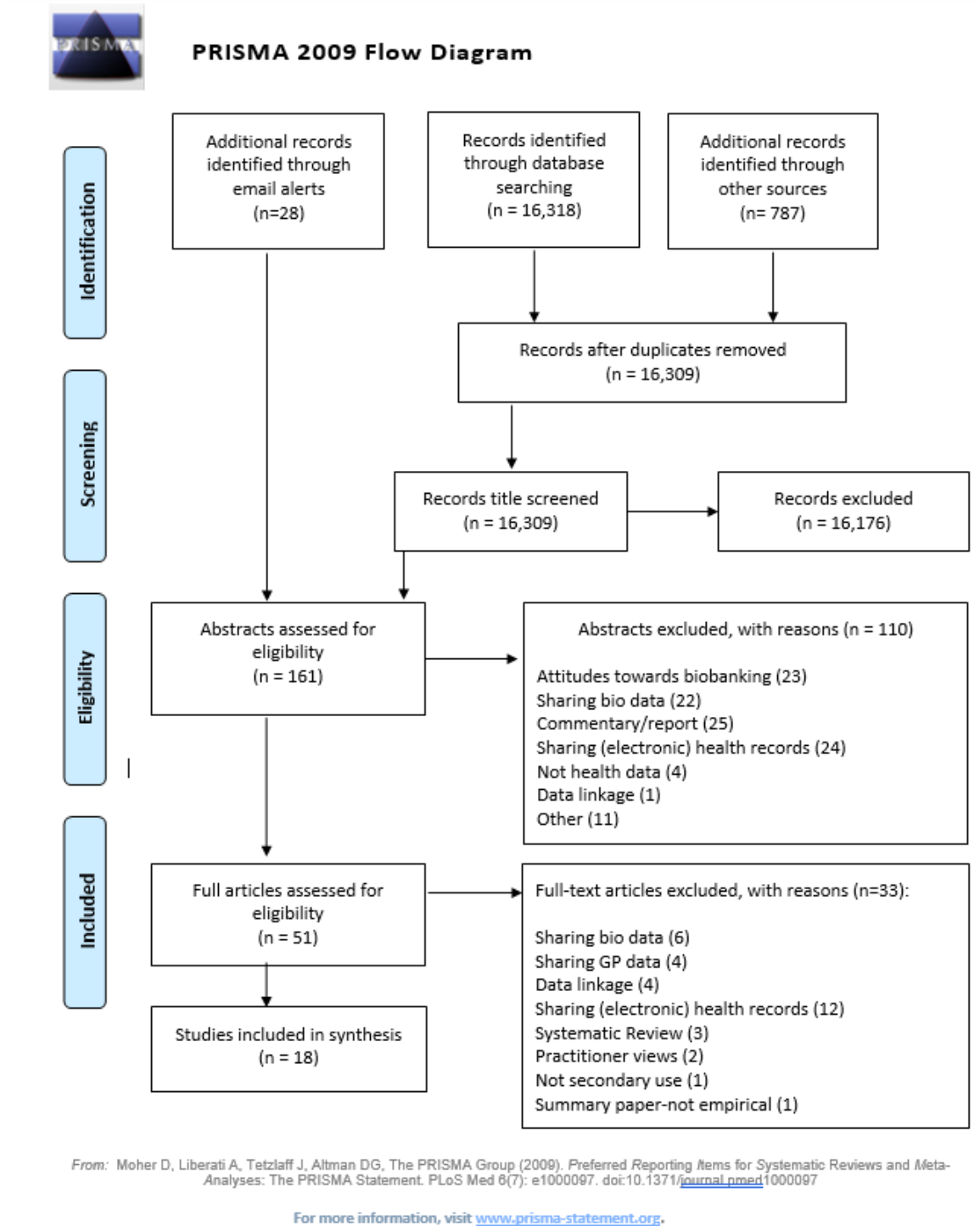


Figure 1: PRISMA flow diagram showing identification and screening process for titles from both database search and email alerts.

2.3.4 Data extraction and quality appraisal

Detailed data were extracted from each included study, including: country of origin, date of research, study design, participant characteristics, study aims, and key themes identified by authors of included papers (Table 1).

The next step was to carry out quality assessment of the included papers. The Centre for Reviews and Dissemination (CRD) in York explain that assessing the quality of included studies gives an *"indication of the strength of evidence provided by the review"* or whether the studies are of high enough quality for their results to be "*believed*" (Centre for Reviews and Dissemination (CRD), 2013, p. 33). Petticrew and Roberts break this down further to state that Quality Assessment assesses *"whether the study is representative of the wider population, whether the numbers add up (for a quantitative study), and whether the study was affected by problems or other events that might affect your interpretation of its results"* (Petticrew and Roberts, 2006, p. 125). If the studies score highly enough, then the evidence they provide can be considered robust enough to inform future decisions on treatment or policy.

Because of the subjective nature of quality assessment, each included study was assessed by two reviewers, myself and one of my supervisors (EM), with results compared and discussed to reach consensus where original assessments differed. Even if an article had aspects that were found to be of lower or questionable quality, it was not rejected at this stage, as the article had met all the inclusion criteria and so was deemed to be relevant. The Critical Appraisal Skills Programme Qualitative Appraisal Tool (CASP) (CASP, 2013) was used for qualitative studies (Asai *et al.*, 2002; Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Jao *et al.*, 2015b; Manhas *et al.*, 2015; Merson *et al.*, 2015; Manhas *et al.*, 2016; Mursaleen *et al.*, 2017a; Cheah *et al.*, 2018; Mozersky *et al.*, 2020) and the Best Bets Survey Checklist Quality Assessment Tool (BestBETs, 2012) for studies using quantitative methods (Platt and Kardia, 2015; Mursaleen *et al.*, 2017b; Platt *et al.*, 2017; Manhas *et al.*, 2018; Mello *et al.*, 2018; Shah *et al.*, 2018; Colombo *et al.*, 2019).

The Critical Appraisal Skills Programme (CASP) is part of the Oxford Centre for Triple Value Healthcare, a social enterprise that supports *"dissemination of knowledge, Learning and skills development"* (CASP, 2013). CASP is a checklist of ten questions that help to appraise qualitative research by addressing three broad issues: *"are the results of the review valid? what are the results? And will the results help locally?"* (CASP, 2013). Each question provides hints or prompts to help the reviewer to answer the question, and each question is answered with either a 'yes', 'no' or 'can't tell'. The final question asks the reviewer to determine how valuable the research is. CASP also provide assessment checklists for other types of research such as systematic reviews or case-control studies.

Best Bets (Best Evidence Topics) (BestBETs, 2012) were developed by the Emergency Department at Manchester Royal Infirmary, and "*arose out of a desire to provide brief reviews of the best evidence about specific topics"*. They provide an online resource of critical appraisal checklists for appraising research such as prognosis, screening, and through the checklist used here, surveys. The checklist asks the user to rate the paper, giving a score from one to 10 and then answer thirty free text questions on sections of the paper such as 'design', 'analysis' and 'discussion' and then rate the paper out of ten again afterwards. I chose to answer the questions using 'yes', 'no', 'not applicable' or 'can't tell' as per the CASP assessment. Neither of these assessment methods provide an overall score but prompt the reviewer to consider the paper's quality and usefulness. The CRD do not explicitly recommend the use of scales or scoring systems to measure quality (Centre for Reviews and Dissemination (CRD), 2013).

2.3.5 Data synthesis and analysis

Results sections from included studies were analysed using the process of thematic synthesis as described by Thomas and Harden in their 2008 paper (Thomas and Harden, 2008). Thematic synthesis is sometimes described as borrowing techniques from grounded theory (Hannes and Lockwood, 2011; Guest *et al.*, 2012) in that it uses iterative or inductive analysis to develop themes that explain data in the same way that grounded theory also uses researcher interpretation to identify *"categories and concepts"* from texts to develop theory (Guest *et al.*, 2012, p. 12). Theories are therefore 'grounded' in the data themselves. Guest *et al* explore an approach to thematic synthesis in their book on applied thematic analysis, explaining that thematic synthesis *"shares the systematic yet flexible and inductive qualities of grounded theory"* (Guest *et al.*, 2012, p. 12).

The thematic analysis involved initial identification of codes (or nodes in NVivo) which was supported by NVivo Software Version 10 for the original nine papers, and NVivo Software Version 12 for the nine later identified papers. Analysis was performed using a systematic yet inductive line by line approach, highlighting all relevant or interesting quotes from participants or descriptions from the authors of the original studies to form 'free codes' based upon their *"meaning or content"* (Thomas and Harden, 2008, p. 5) (see Figure 2), or at least upon my understanding of the meaning. This inductive style of analysis is like a *"form of pattern recognition within the data"* (Fereday and Muir-Cochrane, 2006, p. 82) and is often used when performing thematic synthesis (Braun and Clarke, 2006; Thomas and Harden,

2008; Guest *et al.*, 2012) as opposed to the more deductive process of creating a framework for collecting and sorting references to pre-determined or anticipated themes. The inductive identification of codes in text involves "*careful*" reading of the text of an article to identify potential codes or what Fereday, paraphrasing (Boyatzis, 1998), refers to as to as an *"involved recognizing (seeing) an important moment and encoding it (seeing it as something) prior to a process of interpretation"* (Fereday and Muir-Cochrane, 2006, p. 82). Inductive analysis without a pre-defined framework is referred to as "*data-driven*", although it is important to remember that the researcher cannot entirely escape pre-determined notions or beliefs; *"data are not coded in an epistemological vacuum"* (Braun and Clarke, 2006, p. 84). I did not consider quotes and data from non-participants (e.g., where researchers were also interviewed) or any corresponding author description or classification. It was possible that the same sentence was assigned more than one code. The discussion sections of the quantitative papers were also analysed to provide a greater richness of descriptive data.

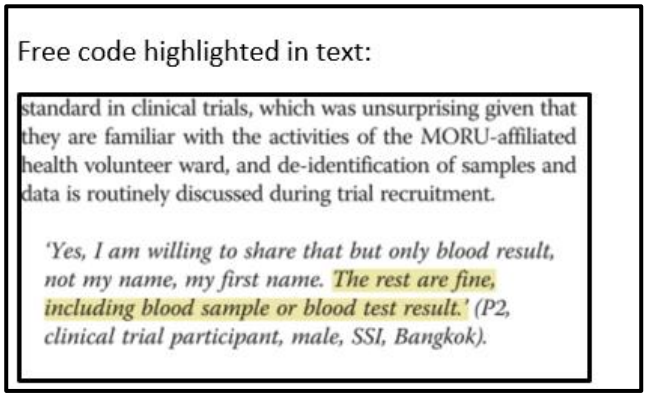


Figure 2- An example of a free code- highlighted text in NVivo.

Free codes thus derived were then amalgamated into descriptive groups, using a hierarchical structure. For the original nine papers in the systematic review publication, this was done by two reviewers, myself and a co-author (EG). This process was repeated until the groups became broad themes, which were reviewed by all supervisors (EM, DNB & TC). Groups were considered suitable to become themes when they contained a large number of similar sub-groups. Grouping codes in NVivo was an evaluative process; the original text was referred to, ensuring that codes were not taken out of their intended context, described by Thomas and Harden as *"grounding a text in the context in which it was constructed"* (Thomas and Harden, 2008, p. 10). An example of this process is demonstrated in Figure 3 below.

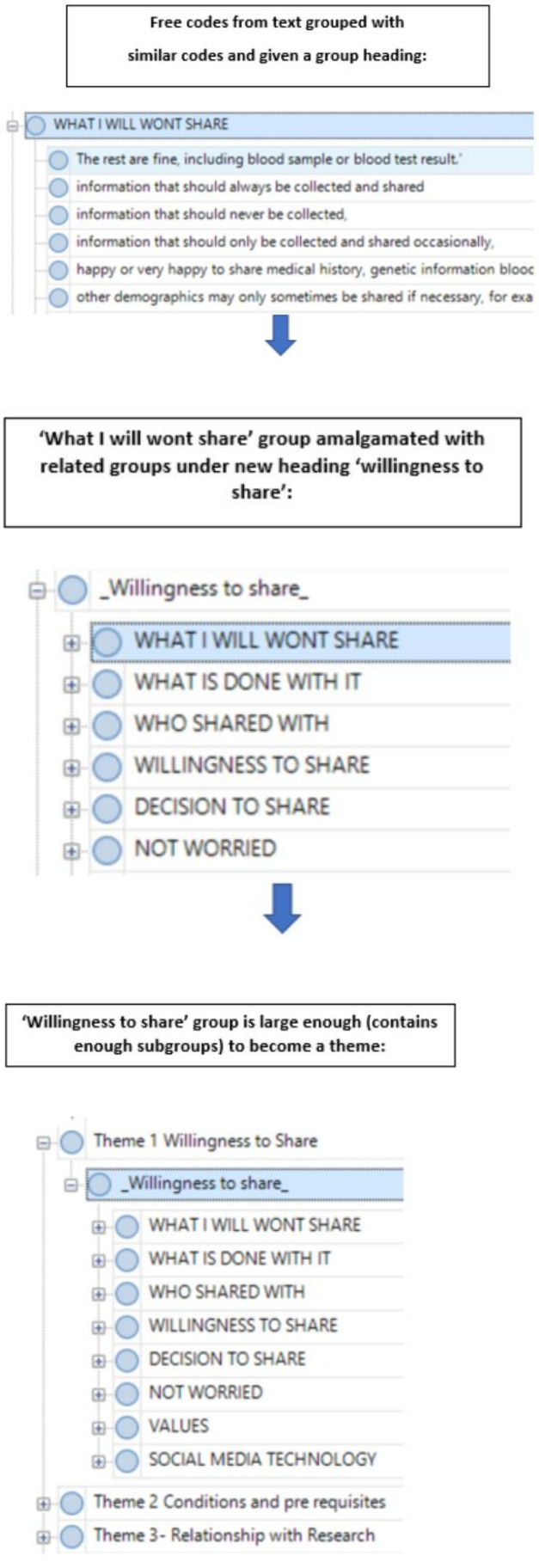


Figure 3- Nvivo screen shots showing evolution of codes to grouped codes to themes.

Although the original systematic review had identified four key themes (benefits of data sharing, fears and harms, data sharing processes and relationship between participants and research), when updating the review and performing analysis on the more recent nine papers, there was no conscious attempt to fit the newly identified codes into the same themes when performing analysis. However, I was conscious that it was possible that my previous findings would influence the codes I chose to highlight, and therefore the themes I eventually arrived at. The nine new papers were analysed in separate files from the original nine and therefore the selected codes (and subsequent themes) from the new papers were not amalgamated or mixed with those from the original nine until I came to the narrative write up of themes.

There was no attempt to produce 'analytical themes' as described by Thomas and Harden, that is using the themes identified both to answer the aims of the review defined at the outset but also to speculate as to how and why the identified themes occur (Thomas and Harden, 2008, p. 7). The purpose of this review was simply to report emerging themes or grouped attitudes towards sharing, not to *explain* the themes on attitudes towards sharing.

2.4 Results

2.4.1 Description of included studies

Of the potentially eligible records identified by database searches (n=16,309), and through subsequent email alerts (n=28), eighteen met the inclusion criteria (Table 1). Nine of these ((Asai et al., 2002; Cheah et al., 2015; Hate et al., 2015; Jao et al., 2015a; Jao et al., 2015b; Manhas et al., 2015; Merson et al., 2015; Platt and Kardia, 2015; Manhas et al., 2016)) were from the original searches (and were included in the published paper) and nine were more recent publications identified from the database email alerts (Mursaleen et al., 2017a; Mursaleen et al., 2017b; Platt et al., 2017; Cheah et al., 2018; Manhas et al., 2018; Mello et al., 2018; Shah et al., 2018; Colombo et al., 2019; Mozersky et al., 2020).

The studies were published between 2002 and 2020, originating from Japan (Asai et al., 2002), Thailand (Cheah et al., 2015; Cheah et al., 2018), Italy (Colombo et al., 2019), India (Hate et al., 2015), Kenya (Jao et al., 2015a; Jao et al., 2015b), Canada (Manhas et al., 2015; Manhas et al., 2016; Manhas et al., 2018), Vietnam (Merson et al., 2015), the UK (Shah et al., 2018), the USA (Platt and Kardia, 2015; Platt et al., 2017; Mello et al., 2018; Mozersky et al., 2020) and the UK/USA combined (Mursaleen et al., 2017a; Mursaleen et al., 2017b).

Eleven studies used qualitative methods, such as focus groups or interviews ((Asai et al., 2002; Hate et al., 2015; Jao et al., 2015a; Jao et al., 2015b; Manhas et al., 2015; Merson et al., 2015; Manhas et al., 2016; Mursaleen et al., 2017a; Cheah et al., 2018; Mozersky et al., 2020)). Seven used quantitative methods such as telephone surveys (Platt and Kardia, 2015; Mursaleen et al., 2017b; Platt et al., 2017; Manhas et al., 2018; Mello et al., 2018; Shah et al., 2018; Colombo et al., 2019), which resulted in fewer direct quotes from participants. Six studies were concerned with research data sharing in low and middle-income countries (Cheah et al., 2015; Hate et al., 2015; Jao et al., 2015a; Jao et al., 2015b; Merson et al., 2015; Cheah et al., 2018). These six studies were part of the same funding award and shared many of the same authors and employed common methods. Three of the included authors (Cheah et al, Manhas et al and Platt et al) had appeared in the original systematic review and had gone on to publish subsequent papers on the same topic which were identified in the email alerts.

Author/title	Country of Research	Study design	Participant characteristics	Aim	Key Themes of study
Asai *et al* (2002)	Japan	Focus group interviews and brief demographic questionnaire with 14 participants.	Lay participants aged 35-55, married with children, with experience or relatives experience of inpatient care during the preceding 5 years. No close family members who were health care professionals.	To explore laypersons' attitudes toward the use of archived (existing) materials such as medical records and biological samples (and to compare them with the attitudes of physicians who are involved in medical research).	• Types of consent • Prerequisites for sharing • Benefits to public • Ownership of medical records • Trust in researchers
Cheah *et al* (2015)	Thailand	Focus group with 7, interview with 1. Topic guides taken from a template developed collaboratively with partners from other sites.	Community members acting as 'community representatives', affiliated with Shoklo Malaria Research Unit where they had been hired as temporary community engagement staff.	To understand attitudes and experiences of relevant stakeholders about what constitutes good data sharing practice.	• Benefits of sharing • Concerns and harms • Suggestions for best practice
Hate *et al* (2015)	India	Focus groups conducted at outreach centres. Attended by field workers as a reassuring presence. Series of scenarios presented that drew on previous contributions to research.	(Employees or) participants in research conducted by Society for Nutrition, Education and Health Action (SNEHA). Participants were familiar with the organisation and its work. 20 female community members.	To identify features of ethical data sharing practice in the context of research involving women and children in informal settlements. Specific objectives were to examine stakeholders' understandings, concerns, and hopes about what would happen to data and their views on what might constitute good data sharing practice; to identify models of data sharing and governance currently in use; to examine contextual considerations affecting data sharing processes; to identify perceived principles of good practice in data sharing; and to consider suitable methods of developing appropriate data sharing processes.	• Benefits of data sharing • Harms of sharing • Barriers to sharing • Obligations and responsibilities • Prerequisites for data sharing • Governance and policy • Broad, middle, and explicit consent.

Author/title	Country of Research	Study design	Participant characteristics	Aim	Key Themes of study
Jao *et al* (2015a)	Kenya	Small group discussions (5-6 people) lasting 3-4 hours. After discussion groups, 3-4 individuals were chosen (reflecting differences in attitude and gender) for interviews lasting 30-45 mins.	A range of stakeholders comprising 30 community members including assistant chiefs (6) and community representatives (24) with relatively low research experience.	A consultation on data sharing, mapping the views and values of diverse stakeholders in a large international research program, the Kenya Medical Research Institute (KEMRI). This paper focuses on views on 'fair processes' in data sharing.	• Types of consent • Informed consent process • Community engagement • Feedback on data sharing process • Oversight for decisions on access to data • Perceived benefits and challenges
Jao *et al* (2015b)	Kenya	Small group discussions (4-6 people) with case study and vignette. Emerging findings noted and used to prompt discussion. After discussion groups, 3-4 individuals were chosen (reflecting differences in attitude and gender) for interviews lasting 30-45 mins	Community representatives- 'typical' community members selected by and from local villages at public meetings to support interactivity for a 3-year period and participate in annual workshops on research related topics.	To report research stakeholders' perceptions of benefits and challenges in sharing data and the emerging importance of trust at individual and institutional levels.	• Importance of data sharing • Challenges and concerns for primary communities • Risks of harms • Fairness to the primary community • Challenges and harms for originating researchers • Misuse of data • Does it matter who's asking?
Manhas *et al* (2015)	Canada	Semi structured interview guide used in focus groups and individual interviews. Recruitment, data collection and analysis continued until data saturation reached.	Maternal and paternal participants in two longitudinal pregnancy cohort research studies. Purposive sampling to identify participants who were fathers and mothers, older and younger than 30, visible minorities and new immigrants. Nineteen people participated in individual interviews and 18 in focus groups (total of 37).	To explore parent perspectives about sharing their own, and their child's non-biological data.	• Altruism has limits • Participants have ongoing privacy concerns • Some participants believe that congruence in values between themselves and research/researchers is important

Author/title	Country of Research	Study design	Participant characteristics	Aim	Key Themes of study
Merson *et al* (2015)	Vietnam	Focus groups with participants and their families.	15 clinical research participants enrolled in observational or cohort studies from northern and southern, rural, and urban centres.	To explore stakeholders' understanding, perceptions, experiences attitudes and concerns about sharing individual level clinical data.	• Views about a novel initiative • Views about acceptable sharing • Trust • Consent
Platt and Kardia (2015)	USA	119 item survey developed to evaluate predictors of trust in the health system, broadly defined as a web of relationships among health care providers, departments of health, insurance systems and researchers. Included 6 trust characteristics included in conceptual model as well as additional questions about trust in specific institutions.	447 members of the general public. 51.5% male aged 18-65 (most aged 26-34). White (76.1%) Black (7.16%), Asian (8.05%), Hispanic (4.70%), Other (3.13%). Most were college or some college educated. 62% non-homeowners. Self-rated health, excellent 18%, very good 40% good 29%, fair 11%, poor 1.6%	To identify characteristics of the general public that predict trust in a health system that includes researchers, health care providers, insurance companies and public health departments. RE Data Sharing in particular: 'our study looks to see whether knowledge impacts trust in data sharing and if so, whether or not it increases support'.	• Knowledge of health information sharing • Privacy concerns • Expectations of benefit
Manhas *et al* (2016)	Canada	Four group (18 participants) and 19 individual interviews.	Maternal and paternal participants in two longitudinal pregnancy cohort research studies. Purposive sampling to identify participants who were fathers and mothers, older and younger than 30, visible minorities and new immigrants.	To examine parent preferences for sharing non-biological data, specifically regarding the consent process.	• Reciprocity: parents want reciprocity among participants, repositories and researchers regarding respect and trust. • Accuracy: parents worry about the interrelationships between validity of the consent processes and secondary data use.

Author/title	Country of Research	Study design	Participant characteristics	Aim	Key Themes of study
Mursaleen et al (2017a) Choices on selective clinical data sharing…	UK/USA	Two focus groups from a total of 43 participants based on findings of survey by Mursaleen et al (below).	Participants were patients with Parkinson's (PwP), predominantly from the USA in attendance at a meeting of patients, advocates, researchers, and care partners.	To characterise attitudes to clinical data sharing among people with Parkinson's disease. Each focus groups addressed 3 questions: Focus group 1: • What data is engaging for people to share? • What data should be collected? • What data is needed? Focus group 2: • How can we inspire people to provide their information? • What personal value is provided by sharing data? • Who should own the shared data?	• Focus group 1 identified data that should never be collected, information that should be collected and shared occasionally or on one-off occasions, and information that should always be collected and shared. • Focus group 2 identified that PwP are more likely to share with assured anonymity and transparency about the use of the data. • Most agreed that data shared by an individual must be owned by an individual.
Mursaleen *et al* (2017b) Attitudes Towards Data Collection…	UK/USA	37 question online survey developed by Parkinson's Movement; an international patient-driven action group created by the Cure Parkinson's Trust.	Paper reports on 310 of 394 patients with Parkinson's disease who completed the 'Sharing Data' section. Roughly even split between males and females. Predominantly UK based but some respondents were from USA and Canada and 17 other countries. Most aged 55-74 years of age. Most respondents saw their neurologist once or twice a year as well as other health professionals.	To establish patient attitudes to ownership of their own medical data and the sharing thereof.	• Focus on collection of symptoms data. • Fewer than half currently shared data- desire to share wasn't necessarily translating into action. • Age was associated with sharing activity. • Sex, medication class and years post diagnosis were not associated with sharing. • Failure of communication suggested as those not sharing were not sure if they were or don't recall being asked.

Author/title	Country of Research	Study design	Participant characteristics	Aim	Key Themes of study
					• Confusion over ownership.
Platt *et al* (2017)	USA	117 item Survey. Knowledge, attitudes, beliefs and trust in relevant institutions, quality of experience, perceived control, and adequacy of policy oversight. Respondents asked 'how true' they thought a statement was.	A nationally representative sample of 1011 respondents. Even split of men and women, 76% white, 9% black, 10% Hispanic. Half of respondents were an employee and 7% were self-employed.	To measure trust in health information sharing in a broadly defined health system (system trust) including health care public health and research and to identify characteristics that predict system trust. Also, to consider any findings in the context of national health initiatives that will expand the scope for data sharing.	• Demographic and psychosocial predictors of system trust. • Implications for precision medicine and learning health systems. • Meaningful transparency in practice: implications for informed consent and the proposed revisions to the common rule. • Building trust: understanding predictors of trust.
Cheah *et al* (2018)	Thailand	Eighteen semi-structured interviews and four focus group discussions with a total of 19 people.	Three groups of participants: 1) clinical trial participants recruited into healthy volunteer studies 2) researchers; and 3) community members with an interest in health research.	Examination of stakeholder perspectives about how best to seek broad consent to sharing data from the Mahidol Oxford Tropical Medicine Research Unit (Thailand). Intended to provide an evidence base for comparison of the merits of different approaches to seeking consent.	• What is it important to know about data sharing? • How much information should be provided about data sharing? • Understandings of data sharing • Suggestions for promoting understanding.
Manhas *et al* (2018)	Canada	An online survey investigating consent preferences for sharing their and their child's non-biological research data.	346 parents participating in two longitudinal birth cohorts: All our Families and Alberta Pregnancy Outcomes and Nutrition.	The final stage of mixed methods research exploring parent's views on privacy, consent, and governance in secondary data use.	• Preferred engagement for consent process • Future communication preferences vs. consent preferences • Consent preferences for child's data

Author/title	Country of Research	Study design	Participant characteristics	Aim	Key Themes of study
Mello *et al* (2018)	USA	10-page structured survey distributed by mail, email or in person distribution accompanied by informed consent and a $40 dollar gift card.	771 current and recent participants from a diverse sample of clinical trials at 3 academic medical centres in the United States. Participants were enrolled or had a child enrolled in a trial within the last 2 years. Included both community-based trials and hospital-based trials.	To find out what participants views are regarding potential harms such as lack of privacy and potential benefits such as maximisation of use of data, in light of journal editors, European Medicines Agency and Food and Drug Administration push for sharing.	• Perceived risks of sharing • Perceived benefits of sharing • Overall support for sharing • Predictors of attitudes
Shah *et al* (2018)	UK	A survey was distributed in Denmark, Sweden, the Netherlands, and the UK. Survey had 3 sections and 24 items covering motivations to take part in and experiences of medical research, opinions on data sharing and socio demographic information.	885 surveys were returned by patients with type two diabetes or those who were at high risk of the disease. Participants were part of one of the DIRECT project studies and were approached by diabetes clinics or university study centres. Participants were aged 18-80 of white European descent and had already consented to data being collected, stored, and shared.	The survey explored data sharing governance to explore the importance of data access governance factors, preferences for which data types may be shared and with whom and who should be involved in managing data access beyond the project.	• Support for data sharing. • Level of happiness for sharing different types of data with different research groups. • Data governance and data access committee preferences.
Colombo *et al* (2019)	Italy	A 22-item online questionnaire in 5 sections delivered via survey monkey.	Questionnaire was sent to 2003 contacts of patient and citizen groups. there were 280 eligible responses.	Italian patient and citizen groups' self-reported knowledge, attitudes, and opinions on IPD sharing, mechanisms for access, advantages, and risks.	• Involvement in clinical research • Awareness of IPD and overall view • Views on access, mechanisms and guarantees for IPD sharing • Risks and advantages of IPD sharing

Author/title	Country of Research	Study design	Participant characteristics	Aim	Key Themes of study
Mozersky *et al* (2020)	USA	Informed consent given online as well as a brief demographic survey delivered by Qualtrics. Interviews then conducted with an interview guide based on a review of the literature on participants views of data sharing.	30 individuals who participated in sensitive (health or health behaviours) qualitative studies. Participants had to be over 18 years of age and participating in at least one qualitative research study. 73% female, 50% white, 50% black or African American and 63% employed, 20% retired.	To explore understanding and concerns regarding data sharing, specifically maintaining confidentiality, secondary analysis, informed consent, and breaching trust.	• Broad support for data sharing. • Concerns about confidentiality and secondary use. • Trust in the research process and in Institutions • Transparency.

Table 1-Characteristics of 18 studies included in the systematic review

2.4.2 Quality appraisal

All studies scored highly in the quality appraisal with many positive 'yes' answers to the questions in both the CASP and BestBets checklists (Table 2 and Table 3).

All but one (Mursaleen *et al.*, 2017a) of the included qualitative studies were rated as 'Quite' (n=3) or 'Very' (n=7) useful with CASP. The CASP question mostly likely to be answered 'no' or 'can't tell' was question 6 'Has the relationship between researcher and participants been adequately considered?'. Six of the eleven qualitative papers provided little or no detail about the relationship between researcher and participant (Asai *et al.*, 2002; Cheah *et al.*, 2015; Hate *et al.*, 2015; Merson *et al.*, 2015; Mursaleen *et al.*, 2017a; Mozersky *et al.*, 2020). Mursaleen *et al* (Mursaleen *et al.*, 2017a) was rated as partially useful overall as in six of the ten CASP questions, the answer given was 'can't tell' or 'partly'.

After assessment with the BestBets checklist, none of the quantitative studies obtained fewer than 20 'yes' answers to the thirty individual questions posed, meaning that there was little missing information. Question 2.9 'What measures were made to contact non-responders?' was judged as 'not applicable' to Mursaleen *et al*, (rather than missing), as the survey link was distributed openly to unknown recipients. The highest rated quantitative study was Manhas *et al* with only three of the 30 questions unanswerable (Manhas *et al.*, 2018). Of the thirty BestBets checklist questions, the most likely to be unanswerable were questions 2.5 '… Have sample size estimates been performed?' and 2.9 'What measures were made to contact non-responders?' with six 'no' or 'can't tell' answers each. Platt and Kardia (2015) was the only study to discuss sample size relative to study objectives; however, they did not explicitly state sample size, response rate and number of non-respondents (Platt and Kardia, 2015). Paper ratings before and after assessment are entirely subjective but are reported in Table 3 to illustrate the assessment process.

Study	Q1 Was there a clear statement of the aims of the research?	Q2 Is a qualitative methodology appropriate?	Q3 Was the research design appropriate to address the aims of the research?	Q4 Was the recruitment strategy appropriate to the aims of the research?	Q5 Was the data collected in a way that addressed the research issue?	Q6 Has the relationship between researcher and participants been adequately considered?	Q7 Have ethical issues been taken into consideration?	Q8 Was the data analysis sufficiently rigorous?	Q9 Is there a clear statement of findings?	Q10 How valuable is the research?
Asai *et al* 2002	Yes	Yes	Yes	Yes	Yes	Can't tell	Yes	Yes	Yes	Quite
Cheah *et al* 2015	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Quite
Hate *et al* 2015	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Very
Jao *et al* 2015a Involving Research Stakeholders in Developing Policy...	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Very
Jao *et al* 2015b Research Stakeholders' Views on Benefits and Challenges...	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Very
Manhas *et al* 2015	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Very
Manhas *et al* 2016	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Very
Merson *et al* 2015	Yes	Yes	Yes	Yes	Yes	Can't tell	Yes	Yes	Yes	Quite
Mursaleen *et al* 2017 Choices on selective clinical data sharing...	Yes	Yes	Yes partly	Yes	Yes partly	Can't tell	Can't tell	Can't tell	Yes	Partially
Cheah *et al* 2018	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Very
Mozersky *et al* 2020	Yes	Yes	Yes	Yes	Yes	Not entirely	Yes	Yes	Yes	Very

Table 2- Quality Appraisal results using CASP

		OBJECTIVES AND HYPOTHESES	DESIGN										MEASUREMENT AND OBSERVATION			
Author/year	**Paper Rating**	**1.1**	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**2.10**	**3.1**	**3.2**	**3.3**	**3.4**
Platt and Kardia 2015	7	yes	yes	Attutides to trust in light of data sharing in general public	yes	yes	yes	yes	yes	can't tell	can't tell	can't tell	yes	yes	yes	yes
Manhas *et al* 2018	8	yes	yes	Participants in birth cohorts	yes	no	can't tell	yes	yes	yes	can't tell	60.80%	yes	yes	yes	yes
Mursaleen et al 2017b Attitudes Towards Data Collection...	8	yes	yes	Survey of people with Parkinson's attitudes to ownership of their own medical data and the sharing thereof.	yes	no	can't tell	yes	yes	can't tell	none-not applicable	can't tell	yes	yes	yes	yes
Platt *et al* 2017	5	yes	yes	Random representative selection of participant panel asked about trust in systems and data sharing	yes	no	can't tell	yes	yes	can't tell	can't tell	52.90%	yes	yes	yes	yes
Mello *et al* 2018	8	no	yes	Attitudes of clinical trial participants to data sharing	yes	yes	can't tell	yes	yes	can't tell	can't tell	79%	yes	yes	yes	yes
Shah *et al* 2018	6	yes	yes	Participants taking part in studies	yes	yes	can't tell	yes	yes	can't tell	none	86% (UK)	yes	yes	yes	yes
Colombo 2019 *et al*	6	yes	yes	Pttitudes of patient and citizen groups to data sharing	yes	no	can't tell	yes	yes	not required	can't tell	15%	yes	yes	yes	yes

	PRESENTATION OF RESULTS			ANALYSIS			DISCUSSION			INTERPRETATION			
Author/year	**4.1**	**4.2**	**4.3**	**5.1**	**5.2**	**5.3**	**6.1**	**6.2**	**6.3**	**7.1**	**7.2**	**7.3**	**Paper Rating now**
Platt and Kardia 2015	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	3/4	yes	6
Manhas *et al* 2018	yes	yes	yes	yes	yes	yes	yes	no	yes	yes	2b	yes	9
Mursaleen et al 2017b Attitudes Towards Data Collection...	yes	yes	yes	yes	yes	yes	yes	no	can't tell	yes	2b	yes	7
Platt *et al* 2017	yes	yes	yes	yes	yes	yes	yes	no	no	yes	2b	yes	6
Mello *et al* 2018	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	2b	yes	7
Shah *et al* 2018	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	2b	yes	8
Colombo 2019 *et al*	yes	yes	yes	yes	yes	can't tell	yes	no	yes	yes	2b	yes	6

Table 3- Quality Appraisal results using Best Bets.

2.5 Themes arising from qualitative analysis

Thematic analysis of the eighteen studies included in this systematic review identified six themes: 1) benefits of research data sharing, 2) fears about data sharing and perceived harms thereof, 3) data sharing processes, 4) relationship between participants and research, 5) willingness to share and 6) conditions and pre-requisites for sharing.

Four of these are also described in the published systematic review (Howe *et al.*, 2018): 1) benefits of data sharing, 2) fears and harms, 3) data sharing processes and 4) relationship between participants and research. The themes 'fears and harms', 'relationship between participants and research' and 'benefits of data sharing' were further supported by additional evidence from papers identified in the email alerts. Two of the themes were identified only when analysing new data from papers identified from the more recent database email alerts: 5) Willingness to share and 6) conditions and pre-requisites. The themes 'conditions and pre-requisites' and 'data sharing processes' exhibit many similarities. For example, both of these themes cover consent preferences and research/data governance issues. There is a more in-depth examination of consent preferences in the newer 'conditions and pre-requisites' and a greater emphasis upon data security and privacy within data governance compared to 'data sharing processes,' which focuses more on participant preferences regarding types of secondary research or researchers.

Each theme is examined in turn below.

2.5.1 Benefits of data sharing

Participants identified the benefits of data sharing, with three main types emerging: benefit to participants or immediate community; benefits to the public more generally; and benefits to science or research.

Most participants wanted to see the benefits of data sharing in their local community, with one participant summarising: *"Data sharing is acceptable if the community benefits…; there is no point in merely writing about issues"* (Hate *et al.*, 2015, p. 244). There should be *"local translational benefits"* (Jao *et al.*, 2015b, p. 8) for *"the community that contributed"* (Cheah *et al.*, 2015, p. 285), particularly if the research in question focussed upon a burden the community faced (Hate *et al.*, 2015).

The *"expectation of benefit"* (Platt and Kardia, 2015, p. 8) from data sharing also extended to the wider public, with phrases such as *"greater good"* (Manhas *et al.*, 2015, p. 90), *"social*

value" (Cheah *et al.*, 2015, p. 285) and *"actually helping people"* (Manhas *et al.*, 2016, p. 6) used in one form or another by research participants. Helping others was a *"dominant theme"* in text comments left by Mello *et al*'s participants (Mello *et al.*, 2018, p. 2206).

This benefit to the public can be reached indirectly by using data sharing to *"accelerate scientific breakthroughs, leading to the development of new treatments or cures"* (Mozersky *et al.*, 2020, p. 17) or to *"get answers to scientific questions faster"* (Mello *et al.*, 2018, p. 2207) by leaving *"no stone unturned"* (Mozersky *et al.*, 2020, p. 17). Scientific benefit was also cited as *"advancement of innovation"* by Colombo and colleagues as one of the main advantages of data sharing, along with reducing waste (of data) and the potential to study side effects of treatments (Colombo *et al.*, 2019, p. 6). The greater the number of researchers who had access to the data, the more likely researchers were to come up with new treatments or cures: *"everybody will know the information and everybody can put their dots together to come up with the solution"* (Mozersky *et al.*, 2020, p. 16). For example, participants with Parkinson's disease thought that collecting data for data sharing could provide a better understanding of Parkinson's or provide personal insights as well as collective insights (Mursaleen *et al.*, 2017b).

Jao and colleagues (Jao *et al.*, 2015b, p. 8) reported that public benefit was sometimes seen as *"satisfied by the involvement of international institutions… such as the World Health Organization",* suggesting that the perception of benefit may be as important as actually experiencing it.

Participants also appreciated the benefits to science and research, explaining that data sharing *"increased the efficiency of research and researcher opportunities"* (Manhas *et al.*, 2015, p. 92), *"generated evidence"* and *"avoided duplication of effort" (Hate et al., 2015, p. 242).* However, benefits were thought by some participants to accrue more to *"scientists in universities and other not-for-profit settings"*, *"physicians"* and *"companies developing medical products"* than to patients themselves (Mello *et al.*, 2018, p. 2208).

Even if data sharing would produce no direct benefit to themselves or their family, participants were still keen to allow data to be shared (Mello *et al.*, 2018). Participants supported sharing to "*enable new analysis*" (Mozersky *et al.*, 2020, p. 18) or verify previous results (Mello *et al.*, 2018; Mozersky *et al.*, 2020), as secondary researchers *"might find something that another researcher overlooked"* (Mozersky *et al.*, 2020, p. 18).

Some participants suggested that data sharing could reduce *"duplication and waste of resources"* (Colombo *et al.*, 2019, p. 8) and had *"cost and efficiency savings"*, saving the researchers' time but also "*saving taxpayer dollars*" as *"I'd rather the dollars get spent once to collect the data, rather than more dollars being spent to collect another set of data when a perfectly valid set of data already exists"* (Mozersky *et al.*, 2020, p. 18). Some participants suggested that responsible sharing could bring efficiencies to participants too or *"keep other researchers from calling me and asking me the same questions"* (Mozersky *et al.*, 2020, p. 18), otherwise referred to as preventing participant burden, in that participants need only be asked once for their data or opinions (Shah et al., 2018; Mozersky et al., 2020).

Participants thought that local researchers should also benefit, and that their *"careers should not be 'overtaken' by others who had made less investment"* (Jao *et al.*, 2015b, p. 9).

2.5.2 Fears and harms

Participants perceived some negative aspects, or potential harms of data sharing, such as risks of being identified, having their data hacked or the study data being misinterpreted.

Mello *et al's* questionnaire asked participants about the potential harms or "risks" of sharing, with 20-26% of participants *"very or somewhat concerned about discrimination, re-identification, and exploitation of data for profit"* (Mello *et al.*, 2018, p. 2204). Participants expressed fear of exploitation, stigmatisation, or repercussions, with some mentioning specific harms that could come to both themselves and their families. For example, harm could be psychological such as *"judgement from others"* or economic such as *"identity theft"* or difficulty obtaining insurance as explained by a participant: *"anything that could potentially compromise your ability to get insurance because it identifies a pre-existing condition…. or reveals, you know, criminal activity or illegal activity…you don't want it getting out there in a way that other people could find out it's you"* (Mozersky *et al.*, 2020, p. 19). Perceived harms ranged from the extreme, such as being reported to social services or an attempted abduction of their child (Manhas *et al.*, 2015), to more mundane concerns, such as third-party contact or telemarketing.

Hacking or information theft was brought up by some participants, *"But, you know, tonight someone could come in and hack the information…"* (Mozersky *et al.*, 2020, p. 20), but not necessarily as a barrier to sharing, more as an accepted potential negative consequence *"You know, with so many hackings and so much information being stolen one way or*

another, I'm not giving it away freely, but I just don't spend a lot of time worrying about it" (Mozersky *et al.*, 2020, p. 20). Information being stolen was one of the most common single *"important"* potential risks in Mello *et al*, but still only selected by 15% of participants (Mello *et al.*, 2018, p. 2204).

Participants wanted to maintain an element of control of their data, highlighting feelings of powerlessness, as there was *"no way for us to know whether or not our personal information is dealt with anonymously" (Asai et al., 2002, p. 6). "Personal information*" was described as "*something that can let people know who you are"* (Manhas *et al.*, 2015, p. 93). Harm was considered more likely to occur if data were shared *"outside the original research team"* even if data were de-identified (Cheah *et al.*, 2015, p. 283), with participants worrying about identification if data were linked with other data or used in ways not initially anticipated. One participant reflected on the need for "*penalties*" for secondary researchers if their data was used in ways *"not affiliated"* with the original research: *"…if they use it for personal gain or a third-party company…"* (Manhas *et al.*, 2015, p. 91).

Concern about being identified or the desire for privacy/confidentiality were referred to in most of the included studies (Asai *et al.*, 2002; Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015b; Manhas *et al.*, 2015; Merson *et al.*, 2015). Some participants talked about the distinction between *"sensitive"* (e.g.: personal details, ethnicity (Cheah *et al.*, 2015, p. 283), HIV status, history of abuse (Hate *et al.*, 2015)), and less sensitive data such as routine demographics (Jao *et al.*, 2015a). As Jao *et al* conclude, the potential sensitivity of data can be related more to its intended use than to the nature of the data itself (Jao *et al.*, 2015a). For Jao *et al*'s participants, sensitive data was more likely to be clinical information regarding a person's illness, diagnosis and management or information on *"sexual orientation and pregnancy status and on socio-economic indicators (such as sanitation, education, and literacy)"* (Jao *et al.*, 2015b, p. 8).

Participants were concerned that data could be *"misused",* (Hate *et al.*, 2015; Jao *et al.*, 2015b; Manhas *et al.*, 2015; Merson *et al.*, 2015) either unintentionally (misinterpretation), or deliberately, to contact participants, or to manipulate data to suit a particular purpose. Some of these participants were informed about data that would and would not be shared through de-identification (Hate *et al.*, 2015; Jao *et al.*, 2015b; Merson *et al.*, 2015) but it is not entirely clear from the papers whether participants were informed about de-identification before or after expressing their concerns about misuse. Misuse such as

unwanted contact should not be possible from an anonymised dataset so perhaps participants did not understand de-identification. Knowing that shared data would be anonymised was reportedly reassuring for Jao *et al*'s participants but did not necessarily convince them that misuse would not occur (Jao *et al.*, 2015b). Some participants were reassured by the reputability of researchers, as this implied governance systems that would reduce the chance of misuse (Hate *et al.*, 2015). *"Misuse"* was therefore about both confidentiality and aligning secondary research with participants' principles.

A specific harm encompassed within *"purposes that trial participants do not approve"* (Colombo *et al.*, 2019, p. 8) was unwanted contact for marketing (Cheah *et al.*, 2018; Mello *et al.*, 2018) or other purposes unrelated to health research such as insurance (Cheah *et al.*, 2018).

Participants in one study identified that bias in research could be a negative consequence of data sharing, with secondary researchers interpreting a qualitative data set in the wrong way, because of their unfamiliarity with it *"a guy looking at it on a piece of paper doesn't have your facial reactions, doesn't have your tone of voice, for instance… so it could be partially biased…"* and *"they may bring their own biases to the study because they weren't the original interviewer and don't know all the ins and outs"* (Mozersky *et al.*, 2020, p. 20). This misunderstanding would then be promoted or spread through the secondary research publication.

Another participant mentioned briefly the training of the researcher in the context of their interpretation of the data *"I would just be concerned about how that second researcher was trained…because if it is different, then obviously, those results are gonna be reviewed differently"* (Mozersky *et al.*, 2020, p. 20). This is supported by approximately 24% of participants in a separate study being "somewhat" or "very concerned" that "*people might use the data to do poor-quality science*" (Mello *et al.*, 2018, p. 2206). To prevent this sort of misinterpretation, participants suggested that their data be shared with researchers who will use it for projects similar to that in which the participant originally took part (Mozersky *et al.*, 2020).

Some participants reported that they would be hesitant to share their data as they were sceptical that it would be used in the right way and were therefore more likely to consent somewhat *"reluctantly"* (Asai *et al.*, 2002; Manhas *et al.*, 2015; Merson *et al.*, 2015). By

contrast, Mozersky *et al* reported that for participants, the risks associated with data sharing were not great enough to stop them agreeing to share their data (Mozersky *et al.*, 2020).

2.5.3 Data sharing processes

Identified barriers to data sharing included their "*novelty" (Jao et al., 2015a, p. 269),* "*limited precedent*" (Hate *et al.*, 2015, p. 243) and practicalities such as the time or work involved to prepare data for sharing (Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Merson *et al.*, 2015). Participants recognised the resources required to implement data sharing, with phrases such as "*resource implications", "funding and capacity building"* (Cheah *et al.*, 2015, p. 284) and "*substantial work"* (Hate *et al.*, 2015, p. 244) used by authors to paraphrase participants' views.

Studies based in low and middle-income countries (Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Jao *et al.*, 2015b; Merson *et al.*, 2015) specifically emphasised community or stakeholder involvement, while participants' desire to be involved in the data sharing process was identified in all studies, as was the desire to be notified when their data was (re)used, and to be informed of the results of studies using their data.

Participants showed varying degrees of understanding of the consent process. Some participants saw the consent process as an informative tool that can play a *"wider educational role" (Jao et al., 2015a, p. 269): "Perhaps you can explain in the consent form… other researchers can access my data to do further research"* (Merson *et al.*, 2015, p. 257).

Seven studies (Asai *et al.*, 2002; Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Merson *et al.*, 2015; Manhas *et al.*, 2016) discussed different levels of consent with participants. Generally, the literature refers to two types of consent, as in Cheah *et al* (Cheah *et al.*, 2015) where a broad consent and a re-consent are described. Broad consent for sharing is given at the consent to the original research study whilst re-consent is generally given by participants on an individual basis when their data are requested for sharing, particularly if the requestee or proposed secondary project is different to those agreed during broad consent (Cheah *et al.*, 2015). Other research has presented additional types of consent to participants, for example, Manhas *et al* explored attitudes towards *"Traditional, opt-in consent", "Broad, one-time consent", "Broad, periodic consent", "Tiered consent"* and *"Opt-out consent"* (Manhas *et al.*, 2015, p. 90).

For some, a broad initial consent would be acceptable, while others wished for *"individual informed consent"* or *"personal permission*" (Asai *et al.*, 2002, pp. 2, 5). Participants evaluated the practicalities of each approach but stated their preference based on ideals of respect and transparency: *"we always like to be asked…I don't think [the project-specific consent model is] a great idea, but I think it would make us feel good*" (Manhas *et al.*, 2016, p. 6). For others it depended on with whom the data would be shared, and they would evaluate on a *"case-by-case basis"* (Cheah *et al.*, 2015, p. 285).

Re-consenting was described in one study as an *"unnecessary inconvenience"* (Jao *et al.*, 2015a, p. 270) and an *"annoyance"* or *"irritation",* (Manhas *et al.*, 2016, p. 8) which risked inviting more questions than if researchers had just shared data anyway. References were made to the practical difficulty of re-consenting participants (Cheah *et al.*, 2015; Jao *et al.*, 2015a; Merson *et al.*, 2015; Manhas *et al.*, 2016).

To be more comfortable with data sharing, participants wanted better data governance or gatekeepers, with processes to store data and manage access requests (Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Merson *et al.*, 2015; Manhas *et al.*, 2016). Research data repositories could act as *"stewards"* for data *(Manhas et al., 2015, p. 94)* perhaps with a committee who could oversee data sharing requests (Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Manhas *et al.*, 2016). A committee would be *"a group trusted to make decisions"* (Jao *et al.*, 2015a, p. 271), ideally with lay or community representatives (Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015), who could reach a consensus, and be held accountable for sharing decisions (Manhas *et al.*, 2015).

Participants identified other conditions that they would like to see in place before they could comfortably agree to share their data, including participants having understood that their data *could* be shared (transparency), risks mitigated, the research being in the public's interest, and the research being congruent with the participants' values (Asai *et al.*, 2002; Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Merson *et al.*, 2015). Researchers did not make any explicit recommendations as to how researchers can ensure that future research aligns with participant values. Some did emphasise the importance of governance and access committees to make decisions on sharing that protected the interests of participants (Hate *et al.*, 2015) and align that research with the interests of participants (Jao *et al.*, 2015a). However, as exhibited in Manhas *et al*, when participants shared their thoughts on industry-based researchers, some participants had

polar opinions *"some parents were reticent to set limits a priori on what secondary research could be conducted with the data, while others felt that certain types of research should be excluded from the start to ensure alignment with parent motives"* (Manhas *et al.*, 2015, p. 94), so it is unclear what decision a committee could make to ensure research aligns with all participant values.

2.5.4 Relationship between participants and research

This theme encompasses the participant's relationship with the research and data sharing process through concepts such as ownership of data, involvement of participants in the research process and how feedback on use of data can be provided to participants.

Some participants exhibited a *"wide range of understanding of data sharing"* while others *"did not clearly understand"* but were nonetheless able to state that they expected that data was de-identified in clinical trials (Cheah *et al.*, 2018, p. 6). Some studies reported that participants were largely unaware that researchers might already be sharing their data (Asai *et al.*, 2002; Hate *et al.*, 2015; Jao *et al.*, 2015a; Jao *et al.*, 2015b; Manhas *et al.*, 2015; Manhas *et al.*, 2016). Despite some participants having at least a partial understanding of data sharing, in one study, one in ten participants who stated that they were willing to share their data did not know whether or not they were currently doing so (Mursaleen *et al.*, 2017b). There was a feeling of *"confusion"* brought about by *"communication failures"* (Mursaleen *et al.*, 2017b, p. 527).

Participants reported that they wanted data sharing to be better publicised, or to be given the option to choose whether or not to share. Nonetheless, when participants were subsequently informed about data sharing, it was largely accepted as a "*necessary sacrifice*" for scientific or medical progress (Manhas *et al.*, 2015, p. 93).

There were then *"high levels of uncertainty about how data might be used once it had been shared"* (Jao *et al.*, 2015b, p. 10), with a desire for transparency regarding the recipient's intentions, and the concept of trust and confidence in research and researchers became apparent in the included papers.

Participants exhibited trust in researchers and institutions to *"appropriately handle"* their data and to choose appropriate projects for secondary use, assuming that *"researchers have the knowledge, skills, and available regulatory or policy guidance to make those decisions"* (Mozersky *et al.*, 2020, p. 21). One participant pointed out that if they were going to take

part in research, they had no choice but to trust the research team *"…I don't really feel like I have a lot of control as to how it will be used. In fact, a lot of times, I don't know. So, again, it comes back to trust… working with an agency that seem reputable and actually who I have a history with"* (Mozersky *et al.*, 2020, p. 21).

The idea that their data could be shared with a secondary researcher prompted participants to consider acceptable types of research or researcher. Although one participant was content with anyone *"[a]s long as it's a qualified researcher"* (Manhas *et al.*, 2015, p. 92), others wanted information about the researchers before agreeing that their data could be shared (Asai *et al.*, 2002), based on the idea that you *"…approve secondary researchers, not their projects"* (Manhas *et al.*, 2015, p. 92).

Most participants agreed that their data should not be used for commercial gain. Participants had less trust in *"drug companies"* or *"insurance companies"* than universities (Mello *et al.*, 2018, p. 2205). *"[T]hird parties"* (Jao *et al.*, 2015a, p. 10; Manhas *et al.*, 2015, p. 92) or *"industry-based researchers*" (Manhas *et al.*, 2015, p. 94) were distrusted because they might use data in a way that was inconsistent with the values of the participant, or attempts might be made to contact them *"for nefarious or unconsented purposes"* (Manhas *et al.*, 2015, p. 95) (e.g., telemarketing). One participant stated that if their data were to be shared with *"a for-profit research group or something, I would want to know and at that point I would actually probably opt out"* (Manhas *et al.*, 2015, p. 94).

Some participants had low levels of trust in procedures for sharing data, with only 42.5% of participants in one study believing that they would not be misled about the use of their data, and just over half (52.3) believing that their data would be used *"responsibly"* (Platt *et al.*, 2017, p. 8). Some participants perceived an overall disparity between researchers' intent to share data responsibly and the degree to which they could be confident in a sharing system's *"integrity and overall trustworthiness"* (Platt *et al.*, 2017, p. 13). A greater degree of trust in research, researchers and repositories was indicative of participants being happy with *"less interaction"* with research teams (Manhas *et al.*, 2018, p. 9).

If participants allowed their data to be shared, researchers should ensure that they make good or proper use of it (Asai *et al.*, 2002; Manhas *et al.*, 2015). It would be *"wrong"* to use the data in a way that the participant is unlikely to have agreed to or understood (Jao *et al.*, 2015a, p. 269; Jao *et al.*, 2015b, p. 13). Participants were placing a great deal of trust in researchers to share their data with appropriate collaborators.

Tied up in the concept of trust was transparency, but not just token transparency such as provision of information, more a *"two-way negotiation of trust"* that can be achieved by engagement, such as being *"responsive to questions from the public"* (Platt *et al.*, 2017, p. 15), which must be present *"from the outset"* (Mursaleen *et al.*, 2017a, p. 31). Colombo *et al* suggest that transparency can be ensured by providing a *"clear definition of responsibilities"* of those sharing data, for example, being open about security and storage, sharing agreements and what will happen in the case of data misuse (Colombo *et al.*, 2019, p. 8). Other methods to demonstrate transparency include being clear about how data will be used (Mursaleen *et al.*, 2017a) and reporting data access requests and results of secondary analysis (Colombo *et al.*, 2019).

The researcher-participant relationship was described as *"socially unequal", a "tacit agreement between the researchers and patients"* (Asai *et al.*, 2002, p. 4) and similar to the *"patient- provider"* relationship (Platt and Kardia, 2015, p. 16), with the researchers indebted to participants (Hate *et al.*, 2015). The relationship with the originating researcher was crucial because it was they who would inform, reassure, and foster a willingness to share. The primary researcher was also the preferred point of contact regarding re-consent *"…just you guys"* (Manhas *et al.*, 2016, p. 7). Participants may have a *"familiarity with physician-researchers"* that brings about a level of trust in research and therefore in data sharing too (Mello *et al.*, 2018, p. 2209).

Participants wanted to be involved in the data sharing process, to have some control over future uses and to receive feedback when their data was used. Shah *et al* refer to participants as *"data donors"* and highlight that most research into participant views of data sharing simply asks participants for *"hypothetical choices"* but not the reasons behind these choices (Shah *et al.*, 2018, p. 11). Shah *et al* involved participants in post-study work to develop a data sharing strategy and propose that ongoing involvement of participants is required if *"research participants are to become integral stakeholders in data sharing governance"* (Shah *et al.*, 2018, pp. 13, 12). One slight caveat when recruiting participants to advise about sharing policies is that participants who are willing to be involved *"typically constitute a small proportion of the people who are eligible for participation and may represent those who are least bothered by data sharing and most enthusiastic about contributing to science"* (Mello *et al.*, 2018, p. 2209).

A further reason to involve participants in the sharing of their data is a concept that was explored by just three of the included papers (two with the same lead author), ownership of data. In Asai *et al* one participant is quoted as saying that data in medical records belonged to the participant themselves and that *"medical professionals and researchers do not have the right to use what belongs to me whenever they want"* (Asai *et al.*, 2002). Medical records specifically are out of the scope of this review, but the sentiment is important. Participants with Parkinson's took part in focus groups and a survey to establish attitudes to sharing their own health data with researchers (Mursaleen *et al.*, 2017a; Mursaleen *et al.*, 2017b). The focus group paper uses the example of data that is collected electronically through apps or devices to monitor Parkinson's symptoms which would then be used in research, so we have to presume this is the data that focus group participants discussed although it is not made explicit. Participants in the focus group agreed that *"data shared by an individual must be owned by that individual" (Mursaleen et al., 2017a)*. This leads to the question of who owns the data once it has been shared with the original research team. Mursaleen *et al* reported that participants were less clear on this when asked in the questionnaire *"who do you think owns your data?*", with approximately 25% *"believing the data was owned by the patients themselves… 38% felt that ownership resided with whomsoever they had chosen to share it with, while a seventh (14.5%) attributed ownership to the platform upon which it was shared*" (Mursaleen *et al.*, 2017b, p. 528). Mursaleen *et al* therefore suggest that ownership of data should be established and communicated at consent stage. It was not explicitly explained in the survey paper what was meant by ownership of data but as identified by the focus group participants it seemed to be linked to permission to share.

Three of the more recently published papers referred to the likelihood of data sharing deterring participants from taking part in research (Mello *et al.*, 2018; Colombo *et al.*, 2019; Mozersky *et al.*, 2020). Just over half (55.8%) of Columbo *et al*'s participants thought that the fact that a study was sharing study data would not deter people from taking part in the first place (Colombo *et al.*, 2019), while Mello *et al* reported that 37% of participants were concerned that data sharing would *"discourage"* others from taking part (Mello *et al.*, 2018, p. 2207). Mozersky's participants were reported that *"risks to confidentiality"* arising from data sharing *"were not great enough"* to deter them from participating (Mozersky *et al.*, 2020, p. 18).

Columbo *et al* (Colombo *et al.*, 2019) and Shah *et al* remind us that the likelihood of participants consenting to share their data was related to their level of trust in organisations or *"gatekeepers"* (Shah *et al.*, 2018, p. 13). Mello *et al* reported that *"low level of trust in people"* indicated a low likelihood to share data (Mello *et al.*, 2018, p. 2207), whilst Platt *et al* point out that trust specifically in health information *systems* for sharing data is critical to their long-term success (Platt *et al.*, 2017). Participants were most likely to feel that the negatives of sharing outweighed the benefits if they also had a low level of trust in *"other people"* generally (Mello *et al.*, 2018, p. 2206).

Finally, several papers referred to researchers providing feedback to participants regarding how their data was shared and used for secondary research. Feedback to participants when their data is used can also be a way to keep the initial consent valid and informed, and provide the right to withdrawal (Shah *et al.*, 2018). To keep track of how data was used Asai *et al*'s participants wanted to be informed *"privately or publicly"* the results of studies that used their data (Asai *et al.*, 2002, p. 5) while Jao *et al's* participants thought that regular feedback on data sharing activities could help *"counter concerns about loss of autonomy and trust"* and provide accountability for researchers from the community whose data was collected (Jao *et al.*, 2015a, p. 272). For participants in Merson *et al* feedback was either fair exchange for their data: *"it must be fair. If you receive my data, you should give me feedback…if sharing is unfair, no one wants to do anything"* (Merson *et al.*, 2015, p. 255) or something that researchers could do to promote the benefits of sharing.

In terms of how this feedback should be delivered, the majority of Manhas *et al*'s participants wanted to hear from researchers once a year via a *"personalised email or a general newsletter"* (Manhas *et al.*, 2018, p. 6), whilst the majority (67%) of Mursaleen *et al*'s believed they should be *"informed when their data is used, most conveniently via email"* (Mursaleen *et al.*, 2017b, p. 526). Manhas *et al* also suggested a password protected account for participants to allow them to access information on how their data was used; this was the second most popular option with participants after email (Manhas *et al.*, 2018). Participants preferred ongoing communication to be specific to the dataset or projects in which their data is held and were *"less interested in general findings arising from the repository's full complement of datasets"* (Manhas *et al.*, 2018, p. 6). It is not reported whether any of the above papers explored with participants the practicalities of feedback.

2.5.5 Willingness to share

Participants exhibited a willingness to share which seemed more pronounced in the more recent papers than in those included in the original systematic review. The willingness to share theme encompassed with whom they would and would not want their data to be shared, and the evident lack of worry regarding sharing of their data.

Several papers (Cheah *et al.*, 2018; Mello *et al.*, 2018; Mozersky *et al.*, 2020) reported participants expressing that they were not worried about sharing their data *"It's something that doesn't have a negative impact to me anyway"* (Cheah *et al.*, 2018, p. 4). This concept of not worrying could be categorised into acceptance of sharing but also not worrying about with whom data was shared or what research was conducted with it *"I'm okay with however they deem the information be used"* (Mozersky *et al.*, 2020, p. 21) as well as acceptance of sharing and not worrying about any potential negative consequences of sharing such a data being stolen *"I don't spend a lot of time worrying about it"* (Mozersky *et al.*, 2020, p. 20), or re-identification, where participants were *"unable to imagine why anyone would care to re-identify them"*(Mozersky *et al.*, 2020, p. 19).

Comments were made explaining that re-identification was not a worry because of the anonymous nature of the data, where participants would be *"like a study number or object"* and *"I don't think anyone would be focusing on just one person" or "I don't think you can pinpoint one particular person"* (Mozersky *et al.*, 2020, p. 19). Overall clinical trial participants believed that the benefits of sharing outweighed the risks (Mello *et al.*, 2018) and that they *"just presumed"* (Mozersky *et al.*, 2020, p. 18) their data would be shared as *"it's the data that is already collected, so I think it's OK to share it"* (Cheah *et al.*, 2018, p. 4).

References to current technology and social media type sharing were made (Cheah *et al.*, 2018; Shah *et al.*, 2018; Mozersky *et al.*, 2020), with participants drawing comparison between data that people freely share on social media or the internet and that which they may choose to share with research teams *"do you people realize how much information you've shared about yourself on Facebook already"* (Mozersky *et al.*, 2020, p. 20).

When it comes to willingness to share data, some participants distinguished between data that they would and would not be comfortable sharing. Mursaleen *et al's* focus group participants identified data they thought should always be collected and shared, data that should be collected and shared occasionally and data that should never be collected and

shared (Mursaleen *et al.*, 2017a, p. 30). There was no distinction made in Mursaleen *et al*'s paper between collecting for an original research team and wider sharing with other research teams, the premise of the paper appeared to be collection of data specifically for sharing to improve scientific expertise. Information that participants thought should never be shared were those items that could *"identify them or influence third party decisions"*, for example insurance or employment decisions (Mursaleen *et al.*, 2017a, p. 30). Data that should be collected and shared occasionally included demographic details such as date of diagnosis or employment. Data that should always be collected and shared were symptom and treatment histories. This study was specific to participants with Parkinson's disease, and so categories such as these used by Mursaleen *et al* may be different for participants of other types of study.

Shah *et al* (Shah *et al.*, 2018) asked participants in diabetes studies about levels of happiness to share and reported that 85% of participants were *"happy"* or *"very happy"* to share details of *"medical history, genetic information blood test results and lifestyle information"* (Shah *et al.*, 2018, p. 10). Other participants referred to the anonymous nature of data that they would be sharing and were therefore comfortable with sharing *"only blood result, not my name, my first name. The rest are fine…"* (Cheah *et al.*, 2018, p. 5), and *"if it's done in a way that my individual information is not shared"* (Mozersky *et al.*, 2020, p. 20).

Mozersky *et al* identified that the nature of the study could determine whether or not participants were willing to share *"If it's like my asthma, I wouldn't mind if my information was not de-identified, but if it was maybe about, like, sex or alcohol… I think I probably would still want it de-identified"* (Mozersky *et al.*, 2020, p. 19). If participants were reassured about anonymity, they were more willing to share: "*she told me that the data that will be shared contains no names or any of my identification. So, I told her it is ok. I will give my consent*" (Cheah *et al.*, 2018, p. 5).

What participants would and would not share can also be attributed to with whom the data will be shared and the sort of (secondary) research that will be performed. In terms of with whom the data could be shared, participants discussed sharing with researchers or scientists and universities, the general public and companies or commercial entities. In Mello *et al*'s survey (Mello *et al.*, 2018), found that 93% of participants were *"very"* or *"moderately likely"* to allow their clinical trial data to be shared with scientists in universities and other not-for-profit organizations, and, although there was less trust in them, 82% would still share with

for-profit companies. Shah *et al*'s study of European participants (Denmark, Sweden, The Netherlands, and UK) reported that participants were least likely to want to share data with drug companies and most likely to share with researchers within Europe (Shah *et al.*, 2018). Other participants stated that they were happy to share with *"other researchers rather than the public"* and that the data should remain within the *"research eco-system"* (Mozersky *et al.*, 2020, p. 18). Only two papers mentioned sharing data with students, with mixed views (Hate *et al.*, 2015; Mozersky *et al.*, 2020). Mozersky *et al* reported that participants were happy to share with students "*for training purposes*" (Mozersky *et al.*, 2020, p. 18) but Hate *et al's* respondents were more divided with some saying that their data could be shared with students, and others suggesting that students should make the effort to collect primary data for their own education (Hate *et al.*, 2015).

Reported reasons to doubt secondary sharing included the potential for misinterpretation of data by secondary researchers *"…I'm not sure if they would get the information correct, so I'm not sure if I agree with that one"* (Mozersky *et al.*, 2020, p. 20).

Closely interlinked with whom the data is shared, is the purpose for which it is being shared. Understanding exactly what the data would be used for was a motivation to share for Mursaleen *et al's* survey participants (Mursaleen *et al.*, 2017b), and understanding what it was to be used for was perceived to ensure that it is not utilised for a project that they would not approve of (Mello *et al.*, 2018; Colombo *et al.*, 2019) or *"anything that they're not supposed to do"* (Mozersky *et al.*, 2020, p. 20).

Projects that participants are most likely to approve of are those which are *"generally similar to its original purpose"* (Mozersky *et al.*, 2020, p. 20), with participants *"most enthusiastic about contributing to science"* (Mello *et al.*, 2018, p. 2209). Projects that participants are least likely to enthuse about are those with *"marketing purposes"* or *"litigation"* (Mello *et al.*, 2018, p. 2202). Sharing data with a project that participants did not approve of was considered a *"major risk"* (Colombo *et al.*, 2019, p. 8).

Participants therefore suggested that there should be a *"a guideline on what they can and can't do with information"* to ensure that data was only used for a "*specific purpose*" (Mozersky *et al.*, 2020, p. 20).

Despite some participants expressing a preference for data not to be shared for certain types of study, other papers reported that participants did not make such strong

differentiations. Columbo *et al* found that participants did not seem to preclude completely *"re-use of data for research questions having a commercial interest"* (Colombo *et al.*, 2019, p. 7), and Mello et al reported *"no appreciable difference"* between "uses that did and uses that did not benefit the participant directly" (Mello *et al.*, 2018, p. 2206).

2.5.6 Conditions and Pre-Requisites

This theme encompasses participant preferences for data sharing processes such as consent, anonymisation, storage of and access to data.

In some of the included studies, (Cheah *et al.*, 2018; Manhas *et al.*, 2018; Colombo *et al.*, 2019; Mozersky *et al.*, 2020), participants were asked about their consent preferences, for example whether they preferred to give broad consent for any future sharing or would prefer to give separate consent for each potential share of their data. There was no general consensus between all included papers, and no consensus between participants within each paper. Participants' identified pros and cons for both types of consent.

According to Manhas *et al* (Manhas *et al.*, 2018) approximately 55% of participants thought that their consent should be sought before data was anonymised and prepared for sharing. Participants wanted to be informed during the initial consent process *"it's nice to know upfront…so that I have a choice one way or the other"* (Mozersky *et al.*, 2020, p. 21) but more as a courtesy than an opportunity for refusal: *"…they should always ask me when I'm signing up… and I'll probably always say yes, but that should just be part of the process"* (Mozersky *et al.*, 2020, p. 21).

In terms of what that consent should look like, '*traditional opt-in*' consent was seen by participants as offering them the most control over their data, but was not the preference for consent models, with approximately 47% of parents whose children had taken part in research preferring the *"least engaging opt-out method"* (Manhas *et al.*, 2018, p. 5). Manhas *et al* also reported that consent preferences were consistent with participant's overall communication preferences (Manhas *et al.*, 2018), indicating that the burden of re-contact for each consent is more important or just as important as the desire to control what happens to the data. Cheah *et al* stated that participants should be informed that re-contact will not involve any additional burden based on participant comments such as *"So the consent to data sharing... means that I have to come back here again or just only this time?"* (Cheah *et al.*, 2018, p. 5). Other participants pointed out that being re-contacted and asked

for consent if a researcher wants to share data for which consent to share had not originally been sought was a *"reasonable compromise"* compared to not being asked at all (Mozersky *et al.*, 2020, p. 21). For other participants, *"agreeing to broad use of their data was inherent in agreeing to participate"* anyway (Mello *et al.*, 2018, p. 2209).

Potential for data sharing should be communicated to participants during initial consent to participate in the primary study, ensuring that dialogue around sharing is transparent and that participants can make an *"educated"* decision as to whether to agree to sharing (Mozersky *et al.*, 2020, p. 21). Participants had *"priority topics"* that they thought should be covered during consent, but the volume and type of information to be provided varied by participant group (Cheah *et al.*, 2018, p. 5). Should participants fail to understand the information provided during consent for data sharing, researchers can end up with *"at worst… an unsafe consent"* (Mursaleen *et al.*, 2017b, p. 527). The importance of this is highlighted by Cheah *et al* who found that participants *"had difficulty recalling the information provided about data sharing"* (Cheah *et al.*, 2018, p. 6). Of the five consent types *"spanning high to low levels of engagement"* put to participants[3] by Manhas *et al*, opt-out consent was ranked as being the least informative (Manhas *et al.*, 2018, p. 2).

Participants had other preferences or required certain assurances prior to agreeing to share their data and these were related to data privacy and security. Some participants were unaware of the way in which data was de-identified prior to sharing *"I really don't know a whole lot, only thing is, like, they said, your name is never used, or your personal information is never used"* (Mozersky *et al.*, 2020, p. 21). Other participants seemed to understand the concept of anonymisation *"I don't think anyone would really pay attention to me in particular because I think it would be a group of just, like, numbers"* (Mozersky *et al.*, 2020, p. 19) but also *"spontaneously sought"* (Cheah *et al.*, 2018, p. 5) assurances that they could not be identified if they consented to future sharing or *"assured anonymity"* (Mursaleen *et al.*, 2017b, p. 526), with several references made to deidentification and anonymisation (Mursaleen *et al.*, 2017a; Mursaleen *et al.*, 2017b; Cheah *et al.*, 2018; Shah *et al.*, 2018; Mozersky *et al.*, 2020). Mursaleen *et al's* (Mursaleen *et al.*, 2017a). Participants also identified that unique identifiers should be used instead of names to provide additional assurance of anonymity.

[3] (1) the traditional consent model; (2) broad, periodic consent model; (3) broad, one-time consent model; (4) tiered (or conditional) consent model; and (5) opt-out consent model

Related to anonymity were references to privacy and security, whereby participants wanted to ensure that their data was stored securely *"under lock and key"* (Mozersky *et al.*, 2020, p. 19) and *"processes and mechanisms"* to reduce the risk of re-identification were in place (Colombo *et al.*, 2019, p. 7). Risk of data loss through unsecure systems or *"security breaches"* was identified by participants who thought it important that as much was done as possible by the researchers who held the data *"as long as the agencies are, you know, doing their due diligence and being honest..."* and what happened after that could not be controlled *"But, you know, tonight someone could come in and hack the information…"* (Mozersky *et al.*, 2020, p. 20). Privacy was particularly important for *"sensitive information"* such as that regarding *"children or other family members, sexual behaviour, potentially embarrassing health information, or their opinions about controversial subjects"* (Mozersky *et al.*, 2020, p. 19). That is not to say that participants would not share this type of information, just that participants required it to be properly protected *"my privacy is sacred"* (Mozersky *et al.*, 2020, p. 19). Participants in Cheah *et al* (Cheah *et al.*, 2018) reportedly recognised ways in which privacy was protected during sharing, and a participant in Mozersky *et* al's study suggested that participants could *"filter themselves"* to ensure their privacy *"if the person doesn't wanna answer certain questions or doesn't wanna let people know too much, I guess they could not choose to do so"* (Mozersky *et al.*, 2020, p. 19).

Further to ensuring that data was secure, and privacy was assured, some participants were asked about access to study data. Approximately 40% of Columbo *et al*'s participants preferred broad access, where *"researchers, representatives of patients' and citizens' associations, journalists and others"* could have access to de-identified data (Colombo *et al.*, 2019, p. 7). Other participants thought that those accessing the data should meet certain criteria *"such as having adequate expertise or supervision or maintaining contact with the original researcher to ensure the interpretation of data is appropriate and accurate"* (Mozersky *et al.*, 2020, p. 20) Other studies' participants supported the importance of *"governance factors"* such as data access request being reviewed by experts, (Shah *et al.*, 2018, p. 6). Manhas *et al* summarised that governance issues (access, monitoring access and bodies involved in access) were even more important to participants than privacy issues (Manhas *et al.*, 2018). Finally, despite 40% of respondents being in favour of broad access, Columbo *et al* also reported that participants required processes such as *"access agreements*

and sanctions in case of data misuse, transparency and public disclosure on access requests and results" in order to agree to sharing (Colombo *et al.*, 2019, p. 7).

Along with assured anonymity, knowing who would access study data was a factor that would encourage participants (38%) to share data (Mursaleen *et al.*, 2017b), with any changes to the originally communicated terms communicated to participants via an *"on-going dialogue"* (Mursaleen *et al.*, 2017a, p. 31). Participants suggested on-going monitoring of where the data was shared (Shah *et al.*, 2018).

2.6 Summary

The available literature on participant attitudes towards sharing data from clinical trials or health interventions is still an area of growth, and this review reflects that, with relatively few studies arising from such a broad search criterion. This study identified six themes, which can be applied to policy or practice and tested with further research.

Previous reviews have explored participants' attitudes towards the sharing of biological and health record data, or data linkage (Stone *et al.*, 2005; Chan *et al.*, 2012; da Silva *et al.*, 2012; Shabani *et al.*, 2014; Aitken *et al.*, 2016a). This review identifies similar concerns: participants are open to, and understand the advantages of data sharing, but they lack awareness and have concerns regarding confidentiality, potential data misuse, governance, and commercial data use.

Participants in the included studies wanted appropriate data protection, and they identified processes that they thought could be modified to promote acceptance of data sharing. Some evidence regarding the effects of data sharing on agreement to participate in research in the first place, only became apparent in papers published more recently (Mello *et al.*, 2018; Colombo *et al.*, 2019; Mozersky *et al.*, 2020).

Chapter 3 Grey Literature Review

3.1 Introduction

This chapter details the process of conducting the scoping review of grey literature from the search through to data extraction and reporting of the results. The relevant guidance is categorised into four main topic areas (guidance on consent, storage, access to data, and types of sharing - four key topics which ran through the themes of the systematic literature review- reported in Chapter 2), and then summarised.

3.2 Background

In light of recommendations that have emerged over the past decade from journals and funders (PLOS., 2014; Institute of Medicine (IOM), 2015; Loder and Groves, 2015; Taichman *et al.*, 2016) that research data be made available for sharing for further research, and collaborative efforts from both funders and research organisations (Walport and Brest, 2011) (HEFCE *et al.*, 2016), there has been an observable trend in the inclusion of explicit sections relating to data sharing or data management in research funder guidance. There has also been an emergence of repositories in which to deposit and store research data for sharing such as the UK Data Archive (UK Data Archive, 2015) or Clinical Trials Data Request (ClinicalStudyDataRequest.com, 2020), with accompanying resources on preparation of data for sharing. There are now even registries built to list and search data repositories (Mendeley, 2020; re3data.org, 2020). Increasingly, data sharing is *"the expected norm"* (Institute of Medicine (IOM), 2015, p. 80).

The advantages of data sharing, such as advancements in science or the speed at which new treatments can be identified, have been well publicised (Walport and Brest, 2011; Institute of Medicine (IOM), 2015; Taichman *et al.*, 2016) and are referred to in the introductory chapter of this thesis (Chapter 1. Background and Introduction, section 1.4).

However, although there is a well-established culture of data sharing in the genetic and genomic communities, data sharing is less ingrained in public health and epidemiological research (Walport and Brest, 2011).

To achieve data sharing, once a research project has been completed, necessitates planning at the outset of studies, with attention paid to consent, storage, anonymisation and future access to data, and these are the areas at which most guidance documents for researchers are targeted.

3.3 Objective

The aim of this grey literature review was to identify and summarise funder stipulations, policy, and guidance documents on best practice for data sharing, and other relevant recommendations for research data sharing in a clinical research or public health research setting. It is specifically focussed on four aspects of data sharing (consent, storage, access and sharing type) identified through the thematic analysis of research participants' views identified in the systematic review (Chapter 2). It was anticipated that these topics, taken from the theme 'conditions and pre-requisites ', were more tangible in nature than those identified in themes such as (for example) 'relationship between participants and research', and so would have practical guidance associated with them. These four topics did however weave their way throughout the systematic review, even in the more abstract themes/subthemes such as the trust between participants and researchers. The specific guidance extracted and summarised within each of the four topic areas was as follows:

- Consent (types of consent, what should be in the consent form);
- Storage (incorporating anonymisation and security);
- Access (including access types, requests for data); and
- Type of sharing (including commercial organisations, trust).

The systematic review (Chapter 2) also found that participants were able to recognise the benefits of sharing, although they expressed fears and perceived potential harms associated with sharing, and that they recognised the importance of the relationship and trust between the researcher and the participant (Howe *et al.*, 2018). It was not anticipated that these themes would receive as much attention in any guidance type documentation; for this reason, whilst they were not ignored if found, they were not the main focus of this grey literature search and analysis.

When referring to trial or study 'data', this review refers to the health data and associated demographic data gathered as part of a trial, longitudinal study or (public) health research study or intervention. These data will typically be information which is collected directly from the participant either via questionnaires or medical tests but does not exclude any other (secondary) data that may be collected as part of the study with the participant's permission, e.g.: from medical, educational, or health and social care records.

Suggestions for best practice identified in this grey literature review are compared, in the discussion chapter of the thesis (Chapter 6), with the attitudes of study participants or members of the public, gained through the questionnaire survey (Chapter 5) and the findings of the systematic review of earlier empirical studies (Chapter 2), supported by the results of the scoping focus group (Chapter 4).

3.4 Materials and Methods

Scoping reviews are designed to "*rapidly map the key concepts underpinning a research area"* (Hidalgo-Landa *et al.*, 2011, p. 46). The current review was carried out following the principle that literature is identified and data are collected and summarised in a structured and reproducible way, but that no quality assessment takes place, and no theories are developed; the data are presented but not necessarily explored for meaning.

For extraction and analysis of relevant literature, guidance was sought from Arksey and O'Malley's recommendations on scoping reviews (Arksey and O'Malley, 2005) and Levac *et al's* paper which builds upon the work of Arksey and O'Malley (Levac *et al.*, 2010) so that the review could be conducted as systematically and transparently as possible.

Newcastle University Library provide 'A guide to useful grey literature sites for researchers in medical sciences' along with the facility to search for grey literature. This was used to identify grey literature along with Newcastle University's Library services webpage on 'Data Curation' which provided access to resources from Research Councils UK and the Digital Curation Centre. Source material for this review were also identified through searching the online European grey literature database 'OpenGrey', which holds 700,000 bibliographic records, for any documents related to data sharing. Web (Google) searches, accessing documents already held (serendipitously obtained) and searches of the websites of key organisations such as the Digital Curation Centre (DCC, 2015), Research Councils UK (now UKRI; links identified via Newcastle University's Library services webpage on 'Data Curation') and UK Data Archive (UK Data Archive, 2015), and funders such as NIHR and Cancer Research UK (CRUK), were also used to identify relevant material. In addition, many documents on data sharing had come to my attention throughout the course of the PhD or through screening of articles for inclusion in the systematic review. Occasionally colleagues or supervisors would send me links to potentially relevant documents via email. Citation and bibliography searches of relevant documents were also conducted.

3.4.1 Search Terms

In 2015 terms related to 'data sharing' and 'policy' were searched for in the aforementioned resources. Searches were repeated in August 2017 and again in August 2019, but contrary to the protocol for this element of the PhD I decided to search only for the term 'data sharing'. By 2017 it was felt that the phrase was well established, and it was therefore unlikely that suitable or relevant literature would be missed by not also using synonyms in the search. A more focused search was also deemed appropriate in reducing the volume of irrelevant guidance that would otherwise need to be screened. After the introduction of GDPR (Information Commissioners Office, 2019b) in May 2018, an interim web search was made to determine whether previously saved eligible documents had been updated to reflect the new guidance. A final search of the literature was made in 2021 prior to thesis completion.

3.4.2 Eligibility

Published (predominantly online) documents detailing data sharing were sought. Some documents might have included just one chapter or section on data sharing alongside guidance on other topics, but that chapter would be considered eligible. The process was partly iterative with few strict criteria for initial inclusion, and eligibility was determined on a case-by-case basis, with decisions recorded for transparency. There was a preference for documents dealing with sharing of data in clinical trials, (public) health research or longitudinal studies, but documents detailing sharing of other types of research data were considered. There was found to be some overlap between documents that set out an organisation's data sharing policy and those that provided guidance on how to implement that policy. The focus was intended to be upon policies or guidance for implementing data sharing, and therefore some well-known statements or editorials on the desirability of and rationale for data sharing, such as the ICMJE statement on data sharing for clinical trials (Taichman *et al.*, 2016), were considered to be out of scope for this review.

The search focused on UK guidance or policy documents published from 1998 onwards, with reference to the date of enactment into UK law of the Data Protection Act (UK Parliament, 1998). It was anticipated that this might be when explicit 'data sharing' guidance was more likely to start appearing, and any pre-existing documents updated to reflect the new legislation. A cursory search of the literature in preparation of this review revealed that, for example, The Medical Research Council published their Policy on Research Data Sharing in 2005 (Medical Research Council, 2016), (and subsequently revised it in 2011 and 2016); the

Wellcome Trust published their first statement in 2007 (with reference to a Fort Lauderdale meeting in 2003) and revised it in 2010 (The Wellcome Trust, 2010); the Information Commissioners Office 'Data Sharing Code of Practice' was published in 2011 (Information Commissioners Office, 2011). Therefore, it was not considered necessary to search earlier than 1998. Eligibility criteria should also be determined according to the time, budget and personnel resources available to the researcher (Arksey and O'Malley, 2005, p. 23), and this was a further argument in favour of a cut-off of 1998.

There were also instrumental reasons for narrowing the inclusion criteria of this literature review. It was deemed necessary to consult only UK literature for this scoping review as it was the intention that the findings would inform policies and practices within the Newcastle Clinical Trials Unit and the Population Health Sciences Institute at Newcastle University, a UK institution, and across the wider community of UKCRC registered clinical trials units. Since data protection legislation and codes of confidentiality and research ethics vary from country to country, UK national guidance was felt to be more relevant in this context. In addition, this review is about summarising the current body of guidance regarding data sharing and not specifically about mapping the growth of data sharing documents over time or across nations.

Once the search had been exhausted (no more relevant documents returned or the same documents returned from different searches), the eligibility criteria were applied to the documents that had been identified and retrieved. Documents were eligible for inclusion in this review if:

- They documented detailed existing requirements of an organisation for data sharing (policy) OR provided suggestions for best practice;
- The publication language was English;
- The publication date was 1998 or later;
- They were applicable to public health, longitudinal or clinical trial research; &
- They were authored by a UK organisation.

3.4.3 Selection Process

The documents identified in the search were downloaded in full (if available electronically), unless it was immediately obvious that they were not eligible. Results were managed in electronic folders and then assessed against the eligibility criteria. This involved reading the

documents contents page, appendix, or skim reading the entire document, depending on the type of document and content. Results were then de-duplicated. When results were found to be eligible for inclusion, references were saved in an Endnote library.

3.4.4 Summarising and reporting the results

All documents identified as meeting the inclusion criteria (and therefore included in this review) were briefly summarised in a table, providing an overview of each document. 'Charting' refers to synthesising and interpreting qualitative data by sifting, charting and sorting it according to key issues and themes (Arksey and O'Malley, 2005, p. 26), similar to data extraction conducted during a systematic review. It is Levac *et al*'s opinion that the process of charting is not well defined, but they suggest that is should be an iterative process (Levac *et al.*, 2010, p. 4). Due to the volume of articles, the chart captured only a very brief summary or key principles of each guidance document.

The chart was designed in Microsoft Word and collected information on the following:

- The type of document;
- The author(s) of the document;
- The year of publication;
- The aim of the article- guidance/best practice/experience;
- The type of data (sharing) explored in the article; &
- The recommendations on best practice for data sharing.

The data collected were then summarised in a narrative account, but unlike a systematic review, there was no attempt to present a specific argument one way or another regarding data sharing. As Arksey and O'Malley (Arksey and O'Malley, 2005, p. 27) point out, a scoping review does not intend to assess the quality of available evidence and so cannot draw robust or generalizable findings. Instead, broad themes are described below, referring back to the original question of best practice in data sharing and any contradicting advice or gaps in the knowledge base will be identified.

3.5 Results

This literature review identified and collated 38 non-duplicate potentially eligible documents, of which 16 contained relevant information on data sharing guidance within the UK and were summarised in a table (see Table 4). The 22 ineligible documents were tabulated, along with reasons for ineligibility, and can be viewed in Appendix C.

Of the 16 summarised documents:

- Six were from funders, outlining data sharing policies that should be adhered to by researchers receiving funding from that organisation (Medical Research Council, 2011; ESRC, 2015; Medical Research Council, 2016; Cancer Research UK, 2017; Medical Research Council, 2017; NIHR, 2019),
- One was a statutory code of practice (Information Commissioners Office, 2020);
- Six were guidance or recommendations for researchers from organisations who facilitate research or provide research guidance (Lowrance, 2002; Corti *et al.*, 2014; Tudur-Smith *et al.*, 2015; HEFCE *et al.*, 2016; Open Research Data Task Force, 2018; UK Research and Innovation, 2018); &
- Three were reports on best practice resulting from workshops or research (The Academy of Medical Sciences, 2013; The Academy of Medical Sciences, 2016; Castell *et al.*, 2018).

The earliest publication was from 2002 and the most recent from 2020, but most were published between 2015 and 2018. Twelve of the 16 documents pre-dated the introduction of GDPR (Information Commissioner's Office, 2018). One of the guidance documents (Corti *et al.*, 2014) was a published book but was also available electronically.

Those that were not eligible for inclusion were rejected because: they were not fully authored or published by UK organisations (n=12); they did not actually contain any detailed guidance on data sharing (n=6) and instead were about data management more generally; and four due to focus solely on biological or healthcare data. See Figure 4, below, for more details.

As an example, the Health Research Authority (HRA) provide links to policies, standards and legislation or guidance on Data Protection and information governance (HRA, 2018), based primarily on GDPR (Information Commissioner's Office, 2018). This guidance covers issues such as legality and transparency in use of data but is not really focussed on data sharing; in fact, data sharing is mentioned only briefly and therefore the HRA guidance is not eligible for or included in this review.

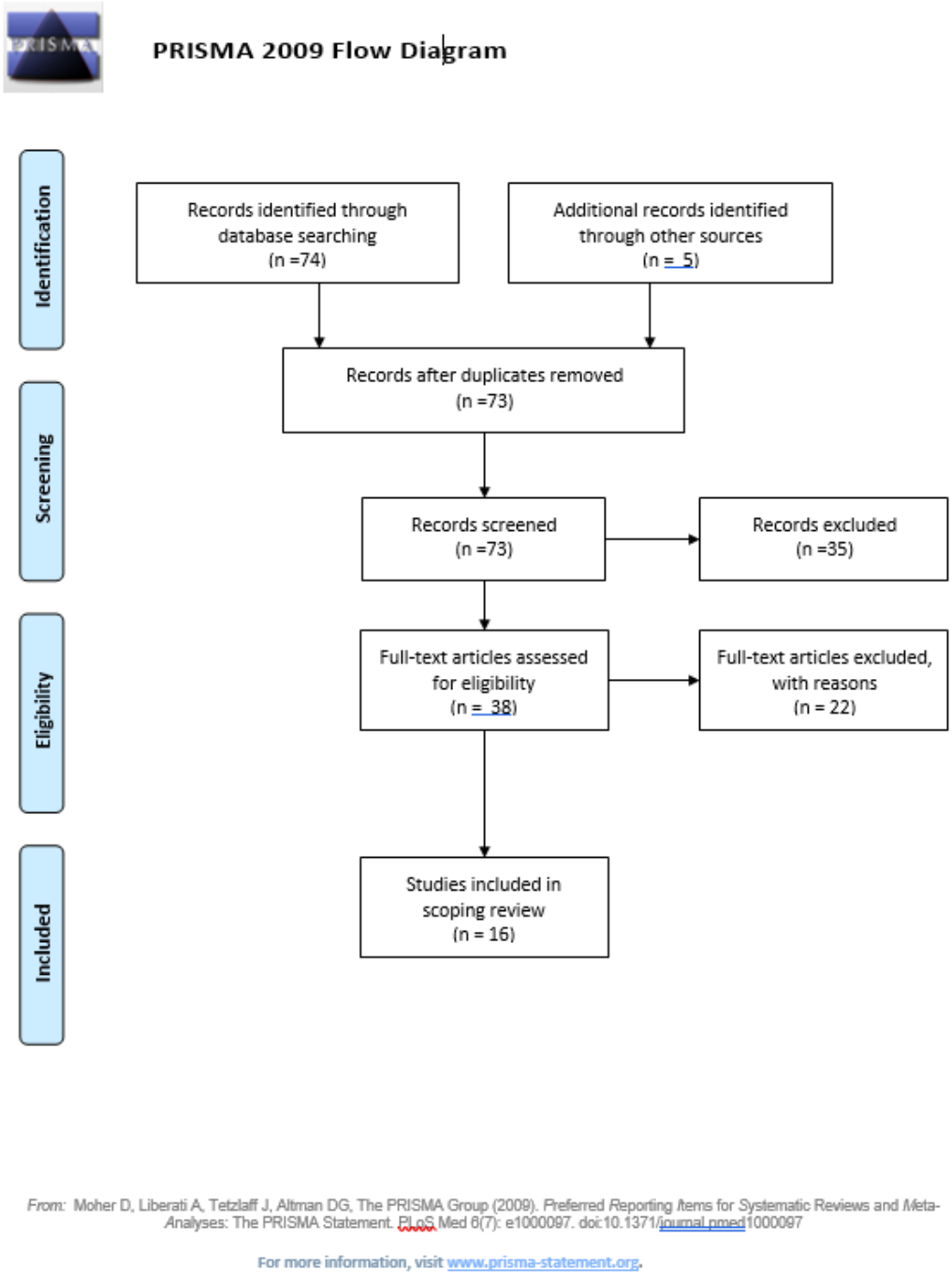


Figure 4- PRISMA flow diagram showing articles identified, screened and excluded

Author/organisation	Title	Date	Purpose/aims of doc
Lowrance on behalf of the Nuffield trust	Learning from Experience-Privacy and the Secondary Use of Data in Health Research.	2002	*"Under what conditions may data not collected specifically for research, such as primary medical data, be re-used for health research without compromising the privacy of the data-subjects?"* (pg. VIII)
Medical Research Council	Policy and Guidance on Sharing of Research Data from Population and Patient Studies.	2011	*"This policy and guidance provides detailed requirements and expectations for individual studies to meet the overarching MRC Policy on Research Data Sharing"*. (pg. 1)
Academy of medical sciences	Clinical trials data sharing: science, privacy, and ethics.	2013	*"The Academy of Medical Sciences brought together experts in clinical trials, ethics, and data privacy, as well as patient representatives, for a dinner discussion on 28 November 2013 to consider what constitutes appropriate access to clinical trial data with a focus on patient level data".* (pg. 1)
UK Data Archive/Corti *et al*	Managing and sharing research data- a guide to good practice.	2014	Book is aimed at researchers to help them remain abreast of changes in legislation relating to the governance of research data or the ethics of research. It can also help reassure professional researchers that their practices are consistent with best practices.
UKCRC, MRC, Wellcome, CRUK, Network of Hubs for Trials Methodology Research (Tudur-Smith *et al*)	Good practice principles for sharing individual participant data from publicly funded clinical trials V1.	2015	Good practice principles for sharing individual participant data.
Academy of Medical Sciences	Summary of a joint workshop to explore the ICMJE proposal on 'Sharing clinical trial data.	2016	The Academy and Wellcome Trust facilitated a workshop to review and discuss the ICMJE's proposal on data sharing (Taichman *et al.*, 2016). This document comprises the notes of a meeting held to discuss the ICMJE proposal on sharing clinical trial data.

Author/organisation	Title	Date	Purpose/aims of doc
Concordat on Open Research Data (Higher Education Funding Council for England (HEFCE), Research Councils UK, Universities UK, Wellcome)	Concordat on Open Research Data.	2016	*"This concordat will help to ensure that the research data gathered and generated by members of the UK research community is made openly available for use by others wherever possible in a manner consistent with relevant legal, ethical, disciplinary, and regulatory frameworks and norms, and with due regard to the costs involved"*. (pg. 1) The Concordat was developed by a UK multi-stakeholder group to provide expectations of best practice reflecting the needs of the research community.
Medical Research Council	Data Sharing Policy.	2016	First published in 2005. The MRC want to *"maximise the research opportunities"* that data provides by ensuring that data are "*properly preserved for sharing"* and informed use beyond the originating research teams. (pg. 3) Their policy on data-sharing builds on the principles developed of the Organisation for Economic Cooperation and Development (OECD).
Cancer Research UK	Cancer Research UK Policy on data sharing and preservation.	2017	*"Cancer Research UK regards it good research practice for all researchers to consider at the research proposal stage how they will manage and share the data they will generate. Therefore, Cancer Research UK requires that applicants applying for funding provide a data management and sharing plan as part of their application".* (pg. 1)
Medical Research Council	MRC ethics series Using information about people in health research V1.0.	2017	*"This guide applies to research using any type of information about people".* *"This guide reflects the current relevant legal framework and will be revised to reflect the new General Data Protection Regulation (GDPR)"*. (pg. 2)

Author/organisation	Title	Date	Purpose/aims of doc
Castell, S., Bukowski, G., Burkitt, R. and Rossington, T. on behalf of the HRA/HTA	Consent to use human tissue and linked health data in health research.	2018	Ipsos mori survey commissioned by HRA and HTA (3 public dialogue workshops and online community). *"A public dialogue for health research authority and human tissue authority"*. (pg. 1)
ESRC	ESRC expectations on Research Data Management and Sharing: ESRC Research Data Policy.	2018	As of March 2015, ERSC grant applicants and grant holders must comply with the council's updated Research Data Policy. *"These principles are aligned with the overarching RCUK Common Principles on Data Policy"*. (pg. 1)
The final report of the Open Research Data Taskforce	Realising the Potential- final report of the open research data taskforce.	2018	*"The Task Force has sought to build on the principles set out in the Concordat on Open Research Data, and to take account of wider moves towards ORD within the international landscape to formulate recommendations"*. (pg. 4)
UK Research and Innovation (UKRI, formerly Research Councils UK)	Guidance on best practice in the management of research data.	2018	Guide to interpreting RCUK Common Principles on Data Policy (2011), which set *"expectations for systematic and routine management and sharing research data".* First published in 2015 but updated in 2018 post GDPR. (pg. 1)
National Institute of Health Research (NIHR)	NIHR Position on the sharing of research data.	May 2019	Statement setting out the NIHR's current potition on sharing of data produced by research that is funded by NIHR. Their policy is in line with the UK Policy Framework for Health and Social Care Research.
ICO (Information Commissioners Office)	Data Sharing Code of Practice.	2020	A practical guide for organisations about steps to share personal data in compliance with data protection legislation.

Table 4- Summary of policy documents included in this review in chronological order

The text sections below detail the guidance from each document (where available) on consent, storage, access, and types of sharing. Sub-headings are given to provide guidance on the observed direction or content of the guidance within each topic area. Some documents provided more detailed guidance than others, and some omitted certain topics (for example, consent) completely. Table 5 identifies which guidance documents covered which areas, although the interpretation of the guidance documents' contents is solely mine. To avoid duplication, not all guidance documents are fully discussed in the text (e.g.: where one echoed the recommendations of another).

The topics for inclusion are as follows:

- Consent (including types of consent, what should be in the consent form);
- Storage (including anonymisation, security);
- Access (including access types, requests for data); &
- Type of sharing (including commercial organisations, trust).

Only two of the guidance documents (Concordat on Open Research Data, 2016; Lowrance, 2002) provided comprehensive coverage of all four areas. Access to data was covered, to some extent at least, in all 16 documents. Guidance or recommendations for consent was given in eleven of the 16 documents. Guidance on storage and type of sharing was provided in twelve documents each. Refer to Table 5 for more details.

Guidance Document (Author, year)	Document contains guidance on:			
	Consent	Storage	Access to Data	Type of Sharing
Lowrance, The Nuffield trust (2002)	Yes	Yes	Yes	Yes
Medical Research Council- Policy and Guidance on Sharing of Research Data from Population and Patient Studies (2011)	Yes	Some aspects	Yes	Some aspects
Academy of medical sciences- Clinical Trials Data Sharing (2013)	Yes	Yes	Yes	Some aspects
UK Data Archive/Corti *et al* (2014)	Yes	Yes	Some aspects	No
ESRC (2015)	Yes	Yes	Some aspects	Yes
Tudur-Smith *et al* (2015)	Yes	Yes	Yes	Some aspects
Academy of Medical Sciences- ICMJE (2016)	No	No	Yes	Yes
Concordat on Open Research Data (2016)	Yes	Yes	Yes	Yes
Medical Research Council- Data Sharing Policy statement (2016)	No	Some aspects	Some aspects	Yes
Cancer Research UK (2017)	No	Yes	Yes	Yes
Medical Research Council-using information about people in health research (2017)	Yes	Yes	Some aspects	Some aspects
Castell *et al* on behalf of the HRA/HTA (2018)	Yes	No	Some aspects	Some aspects
The final report of the Open Research Data Taskforce (2018)	No	No	Yes	No
UKRI (formerly Research Councils UK) (2018)	Some aspects	Some aspects	Yes	Yes
National Institute of Health Research (2019)	No	Some aspects	Some aspects	No
ICO (Information Commissioners Office) (2020)	Yes	No	Some aspects	No

Table 5- Summary of topics covered or omitted in each included guidance document

3.6 Citation linkage

The included guidance documents were also assessed for level of co-citation, in other words the extent to which the documents reference each other and thereby provide consistent recommendations. This process was loosely based upon that of *"Bibliographic Coupling"* first described by Kessler (Kessler, 1963, p. 10) and since adapted for use in various disciplines and also via software designed specifically to map citation and bibliographic links between papers. Figure 5 below displays the level of inter-citation between the included guidance documents. The direction of the arrow is used to indicate which guidance document cites which, with the tail of the arrow showing where the citation was recorded and the point of the arrow showing which guidance was cited. Where a document (for example older guidance such as The MRC Policy and Guidance from 2011) referred to a previous or superseded version of a guidance document, this was recorded but identified with a red arrow. Green arrows indicate that the two documents cited each other.

Some of the included guidance documents were so brief that they did not have any citations at all (Cancer Research UK, 2017; NIHR, 2019), and the remainder provided either a list of references or citations in footnotes. Four of the included guidance documents (Lowrance, 2002; The Academy of Medical Sciences, 2013; The Academy of Medical Sciences, 2016; Castell *et al.*, 2018) did not cite, and nor were they cited by, any of the other included documents, although Lowrance did refer to earlier guidance from the MRC, but not the versions of MRC documents cited here, quite probably due to the earlier publication date of Lowrance (2002).

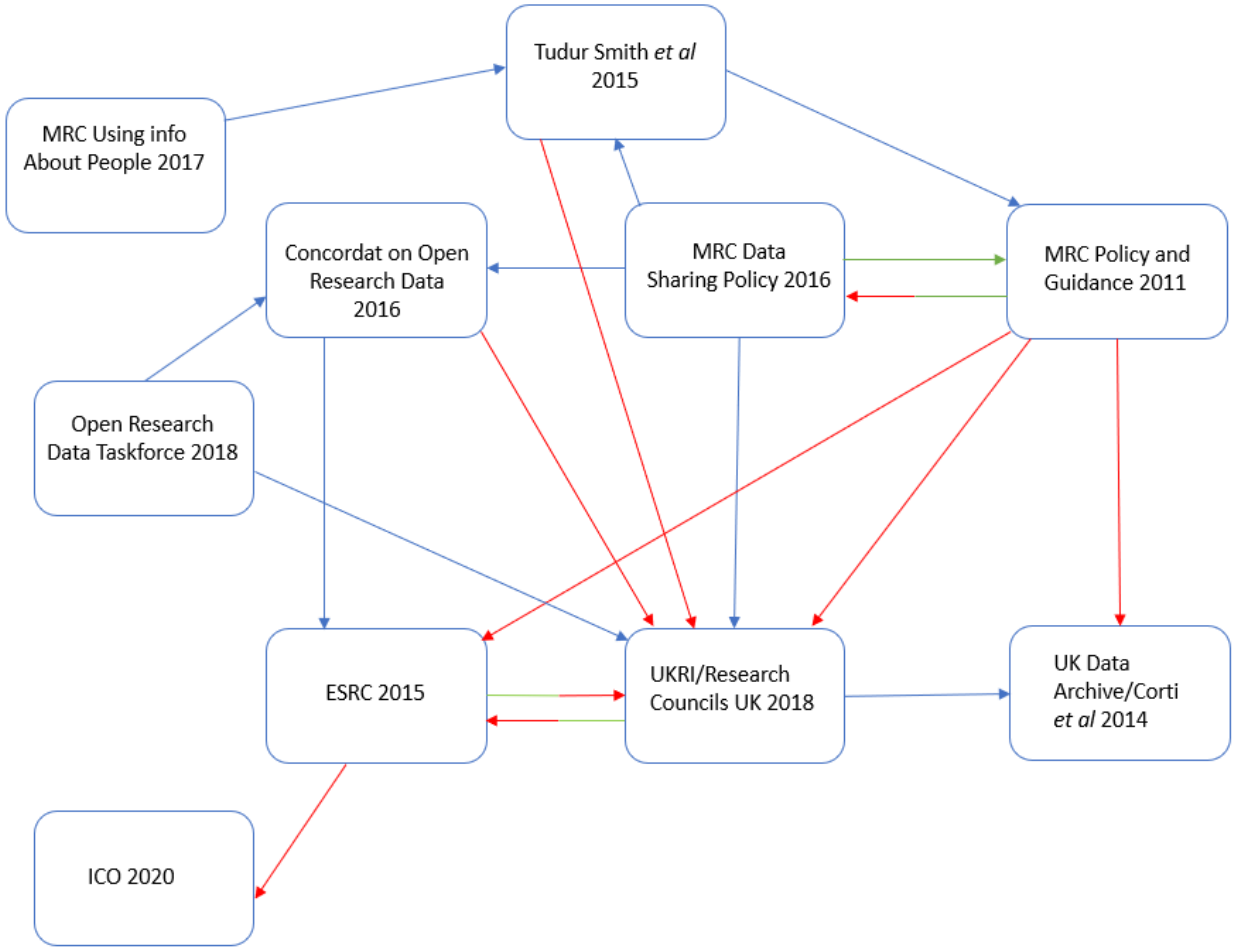
As demonstrated in Figure 5, there are some connections between the included guidance documents. Unsurprisingly the MRC Policy and Guidance (Medical Research Council, 2011) and MRC data sharing policy (Medical Research Council, 2016) cite each other. The MRC guidance 'Using Information about People in Health Research' (Medical Research Council, 2017) seems to be a stand-alone publication, not referring to the other MRC guidance included in this review, although it does cite Tudur-Smith *et al* in the chapter 'Sharing and Publishing'.

The ESRC expectations (ESRC, 2015) are cited by the MRC policy and guidance (Medical Research Council, 2011), The Concordat on Open Research (HEFCE *et al.*, 2016) and UKRI (UK Research and Innovation, 2018). In turn the ESRC expectations only refer to the UKRI. The guidance document most cited by other included guidance documents is the UKRI guidance

on best practice (UK Research and Innovation, 2018), which is logical as UKRI is a combination of research councils, including the ESRC and MRC, and UKRI policy is cited by both. Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) also cite the MRC and UKRI, with the MRC Hubs for Trials Methodology Research being one of the contributors of the Good Practice Principles. Although the Concordat on Open Research data includes UKRI as a member, the Concordat is only referred to by two other documents, the MRC Data Sharing Policy (Medical Research Council, 2016) and the ORDTF (Open Research Data Task Force, 2018).

The Medical Research Council's (MRC) guidance documents 'Using Information about People in Health Research' (Medical Research Council, 2017) and 'Policy and Guidance on Sharing of Research Data from Population and Patient Studies' (Medical Research Council, 2011) sit as accompaniments to their data sharing policy (Medical Research Council, 2016). In this review, for brevity, the 2016 data sharing policy will be referred to as 'Policy', the 2017 document will be referred to as 'guidance' and the 2011 document as 'policy and guidance'. Where no reference is made to the document, the year of publication may be used to distinguish between them.

Key	
Blue arrow	Refers to current version
Red arrow	Refers to previous version
Green arrow	Guidance documents cite each other



No References or citations	No references or citations to included guidance
Cancer Research UK 2017, NIHR 2019.	Lowrance 2002, Academy of Medical Sciences 2013, Academy of Medical Sciences 2016, Castell et al 2018

Figure 5- Citation Linkage diagram

3.7 Guidance on consent

3.7.1 Ethical and Lawful:

In 2002, Lowrance summarised, in the closing remarks of their guidance, that sharing data is expected to pose "*continuing legal and ethical questions"*, such as those around types of consent used (e.g.: implied, detailed or broad, see section 3.7.5) and the rights of a participant to opt out (Lowrance, 2002, p. 70). All of this is wrapped up with implications for participant trust, and potential participant harms.

In response to the release of GDPR, The ICO (Information Commissioners Office) updated their guidance for organisations on *"how to share data fairly and lawfully, and how to meet your accountability obligations"* (Information Commissioners Office, 2020) i.e.: anonymising before sharing or gaining consent to share. This superseded their 2011 sharing policy (Information Commissioners Office, 2011) and was launched with a 'data sharing information hub' with resources for organisations (Information Commissioners Office, 2020)[4]. Predictably, this guidance is heavy on the legal framework around sharing and requires interpretation when applied to a public health or clinical trial setting as it refers primarily to personal data. Indeed, it is not very likely that personal data (defined by the ICO as *"information relating to natural persons who: can be identified or who are identifiable, directly from the information in question; or who can be indirectly identified from that information in combination with other information" (Information Commissioners Office, 2019b)*) would be shared as a result of a clinical trial, health intervention, cohort or longitudinal study. [The 2018 General Data Protection Regulation (GDPR) itself (Information Commissioners Office, 2019b) is also heavily focussed upon the legislative aspects of data protection, rather than data sharing, and so was not eligible for inclusion in this scoping review].

In their 2020 guidance, the ICO stipulate that, in the interests of 'accountability', organisations should check records of consent prior to sharing (Information Commissioners Office, 2020). The lawful basis for sharing data *with* consent is discussed in great detail, but the guidance also goes on to advise that consent is not necessarily always required to share, if there is a good reason or lawful basis for doing so without consent; for example, if the data

[4] https://ico.org.uk/for-organisations/data-sharing-information-hub/

are anonymous, if a task in the public interest is being performed or if sharing is necessary for "*legitimate interests*" (Information Commissioners Office, 2020).

One of the few additional guidance documents to comment on the legal as well as ethical questions regarding data sharing is The Concordat on Open Research Data (HEFCE *et al.*, 2016). This Concordat, authored by the former Higher Education Funding Council for England (HEFCE), UKRI (formerly RCUK), Universities UK and the Wellcome Trust (and since joined by other signatories such as Cancer Research UK), recognises that *"not all data can be open";* where necessary, access to data may need to be managed so as to respect the original consent given and legal, ethical and regulatory frameworks. In fact, *"not all research data can be open"* is one of the key definitions given in the 2016 document (HEFCE *et al.*, 2016, p. 3).

UKRI remind us that consent for data sharing is considered to not just satisfy legal requirements, but to *"maintain and build public trust"* in research (UK Research and Innovation, 2018, p.5), which echoes what Lowrance was referring to in 2002. The NIHR simply state that sharing must be *"consistent with relevant legal, ethical and regulatory frameworks"* (NIHR, 2019, p. 1).

3.7.2 The Ethics of the consent form- where future uses are unknown

Where future uses of data are unclear, the four guidance documents that do refer to unknown future uses, are broadly in agreement that, as far as possible, potential future uses should be communicated in the consent form for consent to be fully ethical. As far back as 2002, in a report that may now be considered dated, Lowrance (Lowrance, 2002, p.x) called for the *"urgent"* development of a new approach to consent, one that differs from the *"classic"* informed consent (which may lack *"ethical validity"*) in that it genuinely provides participants with enough information on which to make an informed choice.

In addition, the point is raised that consent cannot genuinely be 'informed' if we do not know the types of study for which or the organisations with whom data will be shared in the future. This potential absence of 'informed' consent for future sharing was also identified eleven years later by the Academy of Medical Sciences, who point out that the extent to which consent for future studies can be 'informed' is limited when future studies are *"still-to-be-defined"* (The Academy of Medical Sciences, 2013, p.4).

This theme is also picked up by Corti *et al*, who are members of the UK Data Service (UKDS) and make recommendations for consent to enable data sharing in their guidance book on managing and sharing data (Corti *et al.*, 2014). According to them, for consent to be ethically sound, it must be 'valid', in that in must be competent, informed, and voluntary. They therefore argue that for data to be *ethically* shared, consent for future reuse should be planned for and sought at the start point of the original piece of research so that consent (for future sharing) is truly informed. Corti *et al* make recommendations regarding points which *"should"* be included on the consent form, one of which is *"how data may be used for future research or teaching, including any restrictions on that use"* which implies that if researchers are able to anticipate at the consent stage how data may be used, they should do so (Corti *et al.*, 2014, p. 115). In fact, Corti *et al* go on to say that *"possible ways in which their data will be used and by whom"* should be disclosed by giving examples and explaining any limitations on the types of community with whom research data might be shared (Corti *et al.*, 2014, p. 115).

The MRC Hubs for Trials Methodology Research in collaboration with UKCRC (United Kingdom Clinical Research Collaboration), CRUK and the Wellcome Trust (Tudur-Smith, 2015) provide a set of good practice principles to follow for data sharing, which begin at consent. There are no fewer than 15 mentions of consent in this guidance document, although the suggestions are slightly less specific than the detailed guidance on consent in Lowrance (Lowrance, 2002) or Castell *et al* (Castell *et al.*, 2018). For example, the 'good practice principles' document (Tudur-Smith *et al.*, 2015) briefly state that consent needs to be considered from the trial set up through to trial end; a data sharing statement should be included in the consent form and that the level of consent needs to be considered when preparing data for sharing at the end of the study.

In the report from a 2013 Academy of Medical Sciences meeting which included experts in clinical trials, ethics and data privacy, and patient representatives (but not patients) (The Academy of Medical Sciences, 2013), some recommendations regarding data sharing are made. Attendees recognised that there were differing views of what exactly consent for data sharing meant, but there was general consensus that secondary use of data should match the terms of the original consent (The Academy of Medical Sciences, 2013).

This notion of the impossibility of a truly informed consent in face of uncertainty over all potential future uses can be further compounded by the requirement to archive data,

especially if this is in a repository not under the direct control of the original researchers. Nonetheless, while it may not be possible to specify exactly with whom data will be shared once archived, it should be possible to apply restrictions to recipients of archived data and provide examples of likely individuals and organisation with whom data might be shared (Corti *et al.*, 2014, p. 115). Lowrance supports the use of examples; *"it is not impossible to inform broadly and in good faith about possible future uses of data"* (Lowrance, 2002, p.19). Patients could, be told, for example, what their data might be used for in broad terms, e.g.: that it might simply be used to improve diagnosis or to evaluate which treatments were most effective, given that *"a great deal of secondary research proceeds under such general consent, to no apparent detriment"* (Lowrance, 2002, p.19). Tudur-Smith *et al* are in agreement that the presence of an appropriate statement in the consent form regarding sharing is the *"best way to alleviate ethical issues"* at the end of the study, and Tudur-Smith *et al* suggest using HRA wording *"I understand that the information collected about me will be used to support other research in the future and may be shared anonymously with other researchers"* (Tudur-Smith *et al.*, 2015, p.16) (Health Research Authority, 2019).

3.7.3 What should be on the consent form- informed consent versus "information overload":

All guidance documents that consider consent in any depth, (see Table 5) concur that, where known, consent forms or the consent process, should inform participants of proposed future uses of data, with some guidance going further and suggesting that likely projects could also be given as an example (Castell *et al.*, 2018).

Early in his chapter about consent Lowrance uses the words *"non-disagreement"* as a synonym of consent, which at first might seem flippant, but actually well describes the sometimes information-lite version of consent used to ensure that data sharing can go ahead (Lowrance, 2002, p.19).

Amongst organisations who have provided guidance on what consent forms and processes *should* contain are the Health Research Authority (HRA) and Human Tissue Authority (HTA) whose 2018 publication of their Ipsos Mori research or 'dialogue' with participants explored specifically participants' views of consent to link human tissue and health data for research purposes (Castell *et al.*, 2018). Participants were informed that the linked data would be anonymous when used for research purposes, but no detail was given about the linkage process and who would perform this (presumably non-anonymous) task. The aim of this

dialogue was to inform the development of guidance for consent procedures for both the HTA and the HRA (however, neither the HRA or HTA have yet developed their own specific data sharing guidance). Initially though, the report made its own recommendations based upon the finding that participants desired greater transparency and details of the protections in place for their data. A balance needs to be sought between giving participants *"more information which they might not digest" and "the need for informed consent"* or *"information overload"* (Castell *et al.*, 2018, p.3). The six key recommendations of the report are based upon information provision, namely that: information should be provided at consent on: who can access tissue and data; the de-identification process; how donated tissue or data will be used; who can access findings; how tissue donors will be protected and how research findings will be shared. They also state in their report that the consent form should detail about *"access committees"*; how they work and who sits on them (Castell *et al.*, 2018, p.32). Castell *et al* also recommend that more explicit detail is given on the Patient information sheet (PIS) about the risks of sharing, types of researchers who may access the data, how the data are used in research or *"what researchers might be interested in",* and how access committees oversee the process of granting access to data to provide reassurance to participants (Castell *et al.*, 2018, p.46).

MRC guidance agrees with other organisations such as Castell *et al* (Castell *et al.*, 2018), Lowrance (Lowrance, 2002) and Corti *et al* (Corti *et al.*, 2014) that, from the outset, participant *"expectations"* regarding data sharing should be managed through the consent process with *"open information about planned or intended sharing being made at the outset"* (Medical Research Council, 2017, p.22). Although it may not always be clear from the outset with whom precisely data may be shared in the future, an attempt to be as specific as possible at the time of the initial consent is encouraged. Long term plans for sharing, archiving and publishing should be made clear as well in addition to, crucially, the organisation who will be *responsible* for *"keeping it safe"* (Medical Research Council, 2017, p. 16). Castell *et al* take things a bit further and also suggest that examples of typical (secondary) research projects are given on the consent form, as participants had *"such a low base knowledge of health research generally"* (Castell *et al.*, 2018, p. 24) and therefore may not be able to envisage what their data could be used for.

As well as encouraging transparency in the consent process, Castell *et al's* participants themselves also made suggestions for consent, such as providing a glossary of terms, and

making consent forms easy to read by avoiding *"dense language"* (Castell *et al.*, 2018, p.47), advice that is repeated by the ICO; individuals should be informed about what is proposed for their data in a way that is *"accessible and easy to understand"* (Information Commissioners Office, 2020). The MRC (Medical Research Council, 2017) guidance for seeking explicit consent states that the information provided about taking part and regarding how data are stored and used again should be in an "*understandable"* format (Medical Research Council, 2017, p.17). Ensuring the format is understandable can be achieved by testing it using the general public or groups of potential participants, or by following advice from Understanding Patient Data (Understanding Patient Data, 2017).

As well as providing information about sharing, Corti *et al* (Corti *et al.*, 2014) suggest that consent forms should affirm the commitment to confidentiality when sharing. *"Broad but vague"* statements about confidentiality should be avoided and instead replaced with *"specific explanations of how confidentiality will be maintained"* for future analyses, for example by controlling access to the data or through anonymising records. However, the Academy of Medical Sciences suggest that, rather than attempting a *"more exacting standard"* of consent, the focus should be placed on confidentiality via controlled access to data, safe havens, governance and data security measures with *"appropriate sanctions"* in case of data breaches, approaches which they considered would be more ethically sound than a more specific consent process (The Academy of Medical Sciences, 2013). This seems to imply a link between consent and successful anonymisation of data, whereby the consent for future sharing need not be exacting in terms of the types of studies for which future sharing may occur, provided that data are successfully anonymised.

Overall, Corti *et al* recommended that the consent process should take into account uses of the data throughout the lifecycle of a study, from creation through to dissemination, long term preservation and sharing and, at the very least "should not preclude data sharing such as by promising to destroy data unnecessarily" (Corti *et al.*, 2014).

Although researchers such as Castell *et al* (Castell *et al.*, 2018) and Corti *et al (Corti et al., 2014)* recommended that participants should receive information on with whom their data might be shared, there was no indication that in any of the guidance that participants might be able to selectively consent to the various types of organisation.

3.7.4 Where no consent exists:

Where no explicit consent to share data has been obtained, most guidance documents took a pragmatic approach, suggesting that consent be sought, where practical, but that where it was not practical or possible, advancement of research should take precedence over participants' right to be consulted over secondary use of their data.

For example, Lowrance (Lowrance, 2002) suggests that, where explicit permission to share does not already exist, the default stance should be that, where consent can reasonably be sought, it should be obtained prior to sharing for secondary use, with *"urgency, cost, practicality and other factors"* taken into consideration (Lowrance, 2002, p. 20). Whether data are shared with or without explicit consent to do so, it should be assumed (and it should be the case), that *"safeguarding"* and "*independent ethics oversight*" are in place as a minimum (Lowrance, 2002, p. 8).

The view of MRC is also that a researcher must determine whether gaining consent is practicable, or could be made to be so, as research must be based upon "*explicit consent wherever possible"* (Medical Research Council, 2017, p. 16). Just because a study does not legally require consent to proceed to data sharing, does not mean that it should not be sought where practicably possible. The MRC guidance advises that decisions regarding practicality of seeking consent may consider *"attributes of the study population, the research and whether bias may be introduced by seeking consent"* (Medical Research Council, 2017, p. 16).

In Tudur-Smith *et al*, as with Lowrance (Lowrance, 2002) and the MRC (Medical Research Council, 2017), the authors do not entirely preclude sharing where consent is not in place, providing that data are anonymised, and it is not "*reasonably likely*" to lead to the identification of individuals when matched with data available elsewhere. (Tudur-Smith *et al.*, 2015, p. 16).

Other guidance also focusses on the potential for identification of participants and use of anonymisation to enable sharing without consent. The Academy of Medical Sciences suggest that data sharing from historical (and future) studies can be facilitated when data are anonymised in preparation for sharing, so they can be used without ever "needing to retrieve and compare historical consent forms" (The Academy of Medical Sciences, 2013). This of course relies on complete anonymisation (see section 3.8.2 'identifiability to

anonymisation' below for more details of anonymisation) but does remove the need for time consuming re-contact of participants.

The Economic and Social Research Council (ESRC) provide guidance on data sharing for researchers funded by ESRC grants (ESRC, 2015). They recommend that original consent should include permission for future data sharing, and where individual consent cannot be obtained, for example when re-purposing data already collected, data should be *"appropriately"* anonymised or measures should be taken to allow secure access to data to facilitate sharing (ESRC, 2015, p. 7). Every attempt to share data should be made, with lack of consent for sharing not considered an *"acceptable reason"* not to do so. The described approach of using data without explicit consent is described by Lowrance as a *"public interest mandate"* (Lowrance, 2002, p. 20).

The UK Data Service are slightly more cautious, being clear that, before anonymised data can be shared, the consent form and the consent process must be a form of *"active communication"* between the participant and researchers regarding future data uses; the UKDS are very clear that *"consent must never be inferred from a non-response to a communication such as a letter"* (in other words, inferred opt-in) (Corti *et al.*, 2014). By contrast, Lowrance (Lowrance, 2002) sets out the difficulties involved in allowing participants to opt-out of sharing, such as biased data sets, and proposes that the option of opting-out must be addressed in any future data use policy.

The MRC guidance 'using information about people in health research' (Medical Research Council, 2017) provides a useful summary of the legal guidelines regarding consent and summarises the ways in which data can be accessed and shared with or without specific consent, and, for example, when the Data Protection Act 1998 (now superseded by GDPR) did and did not apply to use of data.

It is possible to incorporate consent for any type of future data sharing within the initial consent to take part in a study by using broad consent. The MRC phrase 'broad consent' relates to breadth in both scope and in time, but specificity is encouraged where possible (Medical Research Council, 2017). The MRC Policy and Guidance on 'Data Sharing Requirements for population and patient studies' states that the *"widest range of possible good uses"* of data should be promoted, and that consent should be *"broad and enduring*" (Medical Research Council, 2011, p. 6).

Tudur-Smith *et al* is the only included guidance document to refer specifically to historical data requests for data from studies where no consent for sharing was originally obtained, stating that requests for such data should be dealt with on a "*case by case basis*" (Tudur-Smith *et al.*, 2015, p. 5).

3.7.5 Types of consent:

Types of consent are detailed elsewhere (Chapter 1. Background and Introduction and Chapter 2 Systematic Review) and so will not be covered in any detail here, but the recognition of varying levels or types of consent, such as consent as a *"spectrum"* (The Academy of Medical Sciences, 2013) or terminology such as "Broad" and "Dynamic" consent (Castell *et al.*, 2018, p. 8) were mentioned in several guidance documents, albeit with slightly different phrases used to describe each, as summarised below.

Lowrance concludes The Nuffield Trust guidance (Lowrance, 2002) by presenting three options for consent: 1) use of personal (non-anonymised) data with consent or assent; 2) anonymise data and use it (presumably with consent); or 3) use data without consent under a mandate of public interest. Under option three, data may be anonymised but that will *"depend on the situation"* (Lowrance, 2002). These options are variously presented by other authors such as Corti *et al* (Corti *et al.*, 2014) who state that the consent form should allow participants varying options for consent for participation, publication and sharing. They put forward two options for consent, what Corti *et al* call *"one-off consent"* or *"process consent"* (Corti *et al.*, 2014). One-off consent is blanket consent and covers all aspects of data use including future sharing; however, this is asking participants to agree to potential future uses which may not yet be known. Process consent *"assures active informed consent from participants"* whereby consent for various uses for data can be sought during or after the research is complete (Corti *et al.*, 2014). Castell *et al* refer to this type of consent where researchers re-contact participants as "dynamic" (Castell *et al.*, 2018, p. 8).

However, the Academy of Medical Sciences point out that repeated contact for re-consent prior to each new incident of sharing may not be *"practical"* or *"welcomed"* (The Academy of Medical Sciences, 2013, p. 4). This 'process consent' or 'dynamic' approach does not take into account the potential for loss of contact with participants over the course of the study or the additional cost and administrative burden involved in re-contacting participants and prior to that, the burden of keeping contact details up to date for both parties. The inability to re-contact some participants to obtain consent for sharing leads to potentially biased

samples of data. Once a sample of participants who consented to sharing is defined, an additional burden is ensuring that only their data is actually shared (by excluding from the dataset those who did not consent or were unreachable).

Research Council's UK's, (now encompassed by UK Research and Innovation) document 'Guidance on best practice in the management of research data' (UK Research and Innovation, 2018, p. 6) mentions consent almost in passing, to state that it should be sought, and ideally be *broad consent*, to *"maximise data sharing".*

Castell *et al's* report mentioned the time allocated to consent, and that participants need time to *"digest the information"* (Castell *et al.*, 2018, p. 9) provided as part of the consent process, echoing other HRA guidance that participants should be able to read information provided about the study and sign consent after a *"proportionate"* amount of time (HRA, 2017). Castell *et al* also suggested that provision of REC approved information online would be a useful tool for participants, but that the face-to face consent and opportunity to ask questions about what would happen to their data (and tissue) would continue to be important.

3.7.6 Withdrawal of consent:

Withdrawal of consent is considered only by The MRC (Medical Research Council, 2017) and Castell *et al* (Castell *et al.*, 2018). The MRC require that participants should be made aware that they can withdraw their consent from the original study and that the different types of withdrawal be explained, that is withdrawal from all future analyses or withdrawal from new data collection but allowing use of previous data (Medical Research Council, 2017). The MRC do not explicitly state that participants should be informed that they may withdraw from future data sharing, but this is perhaps implied by the reference to *"future analyses"* (Medical Research Council, 2017, p. 17). In addition, the MRC advise that participants should be informed about the limits of withdrawal, including the point at which it is no longer possible, for example *"post publication"* (Medical Research Council, 2017). The MRC do not state specifically why participants will not be able to withdraw beyond a certain point in the study, but this could be assumed to be due to the difficulty in identifying a single participant's data in an anonymous data set. Opportunities for withdrawal should be offered alongside provision of additional information regarding use of data as and when it is known, for example should the use of data change from that initially agreed, consent should be sought again or as the MRC state, research participants should *"remain appropriately*

informed" (Medical Research Council, 2017, p. 17). The way in which withdrawal will be managed by the study team also needs to be communicated. The study team are required to keep a record of the consent process and the signatures of all involved.

Castell *et al* discovered that participants found that the option to withdraw, as included in the consent form, provided reassurance and *"protection"* from a future of research that may be different to today in terms of *"values, ethics and laws"* and therefore could cause them to re-consider their involvement (Castell *et al.*, 2018, p. 23). Research and science were seen to change rapidly and so having the option to withdraw gave participants a *"sense of control" (Castell et al., 2018, p. 23)*. Participants recommended that, instead of withdrawal being just mentioned, the consent form should specify precisely what is meant by withdrawal, who participants contact to do so; what is done with participants' tissue and data post withdrawal should also be explained in more detail (Castell *et al.*, 2018, p. 38). Participants thought that it would be useful to explain that tissue or data may have already been used prior to participants asking to withdraw. This discussion from Castell *et al* regarding withdrawal was not explicitly stated to be related to data sharing and is found within their general guidance on consent; nonetheless, the points raised are still valid when considering participants right to withdraw from initial research and subsequently from secondary use of their data.

3.7.7 Keeping participants informed of uses:

Interestingly the MRC ask: *"How will you keep participants updated on the research uses, and will your systems support their choice in how data about them is used?"* (Medical Research Council, 2017, p. 17). It is often cited that participants would like to be informed about future uses of their data so that they may either support or decline the use, and so that they may see what research their data has contributed to (Jao *et al.*, 2015a; Manhas *et al.*, 2015; Merson *et al.*, 2015). Althouh it is stipulated as a guidance point, there is, however, no specific guidance from the MRC as to how this should be implemented. Castell *et al* suggest that dynamic consent, and the re-contact that it entails can be used (in theory) as an opportunity for researchers to provide feedback to participants or that *"top level research findings"* can be shared on websites (Castell *et al.*, 2018, p. 46), although this relies on participants actively looking for this information.

3.8 Guidance on storage

All but one of the included guidance documents refer in some way to storage of research data after the original research has ended, for the purposes of access to, and sharing of the data for secondary use. Some of the guidance documents also refer to (an) 'archive' of data, but closer reading reveals that this is simply a term for long-term data storage for the purposes of sharing (Lowrance, 2002; ESRC, 2015; HEFCE *et al.*, 2016), in other words, a repository. However, not all documents provide guidance on storage and there are varying degrees of specificity of the guidance for storage of research data, ranging from simple statements (HEFCE *et al.*, 2016; UK Research and Innovation, 2018), to more specific instructions (Medical Research Council, 2017).

To maximise the discoverability, accessibility, and also the usefulness of research data for future research, it needs to be managed and stored effectively or *"properly curated"* (ESRC, 2015, p. 2), ideally with preparations for storage and archiving being made at the beginning of a study. The Concordat on Open Research Data's principle 6 on data management processes echo these principles, stating that researchers should *"consider how they will manage the data they collect and generate at an early stage of conceptualising their research… A properly considered and appropriate research data management plan should be in place before a specific research project begins so that no data is lost or stored inappropriately"* (HEFCE *et al.*, 2016, p. 14). The ESRC recommend that this early planning and subsequent management of the study data should enable *"data to be exploited to the maximum potential"* for secondary research (ESRC, 2015, p. 3). Research Councils UK state that, in general, data relating to published research should be available for "*at least ten years after publication"* (UK Research and Innovation, 2018, p. 4). Research data should also be stored according to the FAIR principles; findable, accessible, interoperable and re-usable (ESRC, 2015, p. 1; Open Research Data Task Force, 2018, p. 42). The Open Research Data Taskforce (ORDTF) and the ESRC are the only guidance documents to mention these FAIR principles and ORDTF go on to state that data are not always "*discoverable"* enough (Open Research Data Task Force, 2018, p. 42).

In their guidance, the MRC briefly consider the storage of data in the *"longer-term"*, such as archiving so that it can be shared, with research participants being informed of how long the data will be kept and which organisation will be responsible for it (Medical Research Council, 2017, p. 28). In their policy and guidance on sharing data from population health sciences

and population and patient cohorts (Medical Research Council, 2011), the MRC propose the UK Data Archive (https://www.data-archive.ac.uk/) as a suitable repository, but also suggest that researchers may choose a suitable institutional repository.[5]

There is no particular guidance on data management and storage from the Concordat, but instead there is a recommendation that research organisations provide "*guidance to individual researchers on the correct and relevant data management and storage methodologies for that research field"* (HEFCE *et al.*, 2016, p. 14).

Research Councils UK concede that there may be cases in which *"it may not be possible or cost effective to preserve research data"*, and that decisions should be based upon their anticipated "*long-term usefulness*" (UK Research and Innovation, 2018, p. 4).

3.8.1 Storage requirements:

According to the ESRC, a responsible digital repository is a *"digital data repository that takes responsibility for data assets according to the FAIR data principles"* (ESRC, 2015, p. 1).

Lowrance refers to storage of data as *"stewardship"* and questions the length of time that data should be stored in a database (and we can extrapolate this to repositories), arguing that if we are interested in *"societal good"* we may need to store data in a repository indefinitely. (Lowrance, 2002, p. 45). This raises questions of long-term maintenance and security of data, as well as its measured usefulness, which are beyond the scope of this thesis. Responsibility for granting access to data in repositories for secondary research would also need to be managed long-term.

The Medical Research Council seem to consider personal identifiers and the study data as separate entities and state that throughout the life of the study it is the principal investigator's[6] responsibility to ensure that any personal identifiers are separated from the research data as early in the study as possible, and not just at the time of sharing (Medical Research Council, 2017). The principal investigator must also ensure that staff viewing identifiable (personal) data are appropriately trained and understand their responsibility towards participants in terms of *"protecting confidentiality"* (Medical Research Council,

[5] For example, Newcastle University now has a repository for sharing data: https://data.ncl.ac.uk/

[6] MRC consistently refer to 'Principal Investigator' (PI) as responsible for decisions regarding data sharing. Tudur-Smith *et al* use Chief Investigator (CI). In clinical research, the Principal Investigator is responsible for decisions at site, but the Chief Investigator has overall study responsibility. More details on these definitions can be found here: https://www.hra.nhs.uk/planning-and-improving-research/research-planning/roles-and-responsibilities/ accessed 10/06/2021

2017, p. 5). In their guidance 'using information about people in research' it becomes clear that for the MRC, personal identifiers are seen as data used for the process of research, and should not be shared alongside the 'datasets', and should not be requested by secondary researchers unless absolutely necessary for the conduct of the secondary project (Medical Research Council, 2017).

3.8.2 "Identifiability" to anonymisation:

Research data are usually pseudonymised; in other words, all potential identifying information or *"personal data"* such as name and contact details, are removed and replaced with a reference number (Information Commissioners Office, 2019b). Details such as date of birth or ethnicity may remain in the pseudonymised dataset. Reference numbers can usually be traced back to the original personal identifying information by referring to data or files that are stored separately, for as long as those linked records are retained. For this reason, pseudonymisation is more of a *"security measure"* than anonymisation, and pseudonymised data is still subject to GDPR (Information Commissioners Office, 2019b).

By contrast, anonymisation is the process by which research data (which is likely already pseudonymised), is further obscured with potentially identifying information (indirect identifiers, such as age, occupation or location) removed or edited so that participants can no longer be identified through those variables (UK Data Service, 2016). Researchers may need to make trade-offs between complete anonymisation of data and therefore assured anonymity, and utility of data for future analysis. Data that is overly anonymised, to the point of not being able to replicate the original research, loses its utility. It may be necessary for researchers to hold back certain aspects of the data set unless a specific request is made, as discussed in the results section of Keerie *et al* (Keerie *et al.*, 2018).

The guidance from the included documents is largely unified regarding the benefits of anonymisation (although this is sometimes described as pseudonymisation) prior to sharing to ensure that participants cannot be re-identified. Lowrance (Lowrance, 2002, p. 27) describes anonymisation as an "*essential risk-reduction strategy*" to enable sharing whilst the ICO state simply that if objectives can be achieved by sharing anonymous data (rather than personal data) they should be (Information Commissioners Office, 2020). Tudur-Smith *et al* state that "*protecting participant privacy and confidentiality is the over-riding consideration*" when sharing data (Tudur-Smith *et al.*, 2015, p. 17).

Lowrance (Lowrance, 2002) distinguishes between irreversible and reversible anonymisation. Irreversible anonymisation occurs when potentially identifying information is discarded completely or when data sets contain aggregate data, or only aggregate data are shared. Reversible anonymisation (also referred to by Lowrance as Pseudonymisation) simply refers to anonymising data and creating a key (stored separately) that can reverse the process if required (presumably to withdraw a participant). Both of these approaches require judgement calls and potentially "*statistical expertise*" to enable identification of the most appropriate type, or "*acceptable degree*" of anonymisation (Lowrance, 2002, p. 29)- Lowrance does not specify as to which is the most appropriate; presumably, it depends on the type of data to be shared and the type of sharing to occur. There are also judgement calls to be made about the potential later use of the key to reverse anonymisation, and how this should be securely stored (Lowrance, 2002, p. 29).

In their guidance on 'using information about people in health research' the MRC state that *"reducing the identifiability"* of data, and the subsequent risk of identification is *"essential"* to ensure participants' privacy, and can be achieved by anonymisation, pseudonymisation, encryption and restricting access (Medical Research Council, 2017, pp. 19-21). The MRC go on to provide a whole chapter of principles and guidance on anonymisation and pseudonymisation and some tips on how best to reduce "*identifiability*" (Medical Research Council, 2017, pp. 19-21). Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015, p. 23) also provide guidance on anonymisation with a list of 28 items which could be considered identifiable and advise that data should be tested prior to sharing to ensure that it has been anonymised successfully.

According to the MRC guidance, identifiability is a continuous *"grey-scale"* from (fully) anonymous to identifiable, where both the content (identifiers such as name, address) and the context (how it is processed or what it is combined with) of the data determine where on the spectrum the dataset sits (Medical Research Council, 2017, p. 19). The MRC's suggestion to reduce identifiability is to establish at the planning stages of projects whether or not identifiable information is required at all for the purposes of the research. They suggest that access to identifiers should be on a need to see basis, with some emphasis on the role of the principal investigator (sic) to decide on the level of anonymisation or pseudonymisation, and that access to identifiable data should be granted only where appropriate (Medical Research Council, 2017).

The MRC also suggest that if collaborators do not need to see data that are identifiable (which would include pseudonymised datasets), but can work just as effectively with anonymised data, no identifiers should be sent to them; similarly, if the entirety of the data set is not required for secondary analysis, only the relevant variables should be sent (Medical Research Council, 2017).

More specific guidance on conducting the *"craft of anonymising"* can be found in Lowrance (Lowrance, 2002, p. 30), Corti *et al* (Corti *et al.*, 2014) and Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015).

3.8.3 Data security/secure access

The guidance documents which discussed data security were, unsurprisingly, unanimous in the need for anonymity for participants, but it became obvious that there was also a balance to be struck between caution and facilitating access to data, with some guidance detailing the types of controls that would enable secondary researchers to access data securely.

The MRC state that, to keep data "*safe*", controls on data should focus on reducing the identifiability of data, minimising the risk of (re)identification and thereby ensuring data security (Medical Research Council, 2017, p. 25). Research Councils UK advise avoiding "*violation of privacy*" and "*harm to intellectual property*"; presumably the violation of privacy refers to participants and the latter to researchers and/or their host organisations (UK Research and Innovation, 2018, p. 3).

According to the Academy of Medical Sciences, the risk of identification is "*very low*" for "*highly aggregated data*", and therefore data security, and anonymisation should be focussed upon individual level data (The Academy of Medical Sciences, 2013, p. 3). The ESRC also suggest that there are levels of risk and that access to data can depend upon the sensitivity of the data, with *"access Levels"* used to classify data to *"determine the conditions under which access will be permitted"* (ESRC, 2015, p.2).

The Academy of Medical Sciences point out that anonymisation is not necessarily completely secure; nonetheless a balance needs to be struck between level of anonymisation and utility of data, as it is possible that the more "*stringent the anonymization the greater the loss of useful data for secondary analyses*" (The Academy of Medical Sciences, 2013, p. 3).

For data that are potentially too "*sensitive, confidential or potentially disclosive*" the ESRC suggest that access could be via a "*secure access infrastructure*" (ESRC, 2015, p. 5) whereby

data can be accessed in a "*protected virtual environment*" or "*secure access service*" (Corti *et al.*, 2014) – secondary researchers can access and then analyse data on a server located within the host's institution, and once analysis has been completed, the output can be approved for release by the host institution. The Academy of Medical Sciences echo the potential utility of safe havens which employ "*both technical and contractual safeguards*" (The Academy of Medical Sciences, 2013, p. 3). Safe havens as security could also enable a "*layered*" approach to data provision, with "*different types of access for different types of data*" (The Academy of Medical Sciences, 2013, p. 4). Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) also explain the way in which data can be released securely, such as by being uploaded to a repository with restricted access where access can be controlled and removed at any point by the data custodian, for example when access is no longer required, and where data can be viewed but not downloaded.

The ESRC mention briefly the documentation, storage, backup and security in terms of research or academic community standards and "*long-term sustainability*" of the research data (ESRC, 2015, p.3). The MRC say little about storage of data in their data sharing policy (Medical Research Council, 2016, p. 4), but do state that risks, such as *"inappropriate disclosure"* should be managed proportionately, so that opportunities for new uses of data are *"maximised"*. They also provide the caveat that this management of risk and sharing of data must fall within *"regulatory requirements of the law"* (Medical Research Council, 2016, p. 4).

3.9 Guidance on access to data

All included guidance documents referred to access to data to one extent or another. Guidance on access to data centred upon types of access and the processes involved in providing researchers with access to research data, for example reviewing requests for data, preparing data for sharing and recording the process formally in data access agreements.

For the Academy of Medical Sciences (The Academy of Medical Sciences, 2013), open access to individual level trial data raises concerns about privacy and risks of identification, and therefore, measures need to be put in place to control access to data.

One such group who propose measures in the form of a set of principles to *"respect the needs of all parties"* is the Concordat on Open Research Data (HEFCE *et al.*, 2016, p. 4). Their principles intend to establish a set of *"expectations of good practice"* which actually increase

access to research data for secondary research, ultimately for public benefit. They note that *"Extensive statutory and regulatory standards already exist to govern research practice and data access where it is deemed necessary…the concordat does not supersede or replace these but addresses directly the issues related to open research data"* (HEFCE *et al.*, 2016, p. 6). Principle 7 details the commitment researchers should make to ongoing curation and storage of data but does little to further the need to ensure privacy of participants or how access might be controlled.

The Concordat also suggest that there is also a need for more specific guidance in many disciplines to guide researchers and that learned societies may play a key role in developing relevant discipline specific guidance (HEFCE *et al*., 2016).

References to the legal aspects and principles of sharing data, such as the Data Protection Act 1998 , that sit above any individual sponsor or regulatory recommendations made in the included guidance, are also referred to in other guidance documents such as those by the MRC (Medical Research Council, 2017) and Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) and in principle 4 of Research Council's UK guidance where "*individual participants' rights should not be damaged by inappropriate release of data*" (UK Research and Innovation, 2018, p. 5). More recent guidance refers to GDPR (Information Commissioners Office, 2019b) as an overarching principle governing access to personal data. GDPR is mentioned briefly in this context by the MRC (Medical Research Council, 2017), Open Research Data Taskforce (Open Research Data Task Force, 2018) and ICO (Information Commissioners Office, 2020).

3.9.1 Too confidential to share?

Some of the included guidance make reference to data that should not be shared (ESRC, 2015; Tudur-Smith *et al.*, 2015; Medical Research Council, 2017). For the ESRC (ESRC, 2015), not depositing data in a repository is very much the exception for ESRC funded projects; but it is suggested that if the data are considered too confidential to share, the researcher should contact the suggested repository (data service provider) to discuss this at the "*earliest opportunity*", presumably to discuss amendments to access types (ESRC, 2015, p. 2).

Examples of data that would be too sensitive or confidential to share are not provided by the ESRC, but examples given by the MRC, are as defined in the Data Protection Act, including *"sensitive personal data"* relating to matters such as ethnicity, political or religious beliefs, and mental or physical health conditions (Medical Research Council, 2017, p. 31). Other

examples of potentially sensitive data come from Tudur-Smith *et al* in their list of 28 potential patient identifiers in datasets, where examples of *"sensitive data"* include *"illicit drug use or risky behaviour"* (Tudur-Smith *et al.*, 2015, p. 23).

In terms of identifiable data, the MRC stipulate that identifiable information should only be shared if there is a lawful basis to do so, if it is what "*participants would expect and the amount and sensitivity of the data is minimised*" (Medical Research Council, 2017, p. 22). Even where data has been anonymised, the MRC advise that care should still be taken that re-identification could not take place. Again, researchers should refer to Tudur-Smith *et al*'s list of potential identifiers which could inadvertently identify a participant if included in a dataset (Tudur-Smith *et al.*, 2015).

The Open Research Data Task force recognise that whilst promotion of Open Research Data is desired, any "*necessary restrictions on access*" should be "*clearly articulated*", particularly where they relate to the above mentioned sensitive or potentially identifiable data (Open Research Data Task Force, 2018).

3.9.2 Access types:

Descriptions of access types, such as open and controlled, are available from Tudur-Smith *et al* and The Academy of Medical Sciences (The Academy of Medical Sciences, 2013; Tudur-Smith *et al.*, 2014). Generally, open access refers to preparing (anonymising) and placing data in a repository where it may be accessed for secondary use by other researchers or even members of the public who register to access the repository. Controlled access refers to data which may still be placed in a repository, but access is controlled. Requestees will need to meet certain criteria to be able to access the data and access requests may be reviewed by a committee or decision maker. Of all the included guidance, this is best described by Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015).

Most of the included guidance documents take similar approaches to access types, with caution and control, and in some cases a study specific approach to access, advised. However, Research Councils UK call for data to be "*openly available*" with "*as few restrictions as possible*" as well as in a "*timely*" and crucially "*responsible manner*" (UK Research and Innovation, 2018, p. 3).

Other organisations place more emphasis on responsible sharing and access and less emphasis on data being openly available. The ESRC consider the way in which data can be

shared in a responsible manner, lest the research process be "*damaged by inappropriate release of data*" with the application of constraints on data to be considered at the "*initiation of the research process*" (ESRC, 2015, p. 2). According to the ESRC, data can be made available either: *"for re-use free of charge, as open data, safeguarded data or controlled data,* with the access category "*selected to minimise the risk of disclosing personal information"* (ESRC, 2015, p. 3). The distinction between 'safeguarded' and 'controlled' access is, however, not explained in the guidance document.

The MRC policy on data sharing, (Medical Research Council, 2016, p. 4) states that "*Access policies and practices for new and existing MRC-funded data collections must be transparent, equitable, practicable, and provide clear decisions consistent with MRC data sharing policy*". Having an access policy set out at the beginning of the study echoes what the ESRC say about deciding on access type or constraints at study initiation. In their guidance for population and patient cohort studies they had previously stated that a "*simple*" study-specific policy on data sharing which aligns with the MRC's overarching policy (Medical Research Council, 2016) should be available or "*readily discoverable*" by the research community on the study website (Medical Research Council, 2011, p. 3). The idea of a study specific data sharing policy from the MRC mirrors that from the ESRC where *"access category"* is selected as either open, safeguarded or controlled (ESRC, 2015, p. 3).

Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) examine both controlled and open access approaches to data sharing with examples given of both. Tudur-Smith *et al* describe open access as *"anonymised IPD is made available to researchers and the public alike; no approvals are required and there are no limitations or restrictions on the use of the data"* (Tudur-Smith *et al*., 2015, p. 10). The good practice principles do, however, recommend a controlled access approach (*"data requesters have to provide information to support a request for data access"*), backed up by their survey of UK CRC Registered Trials Units, none of whom supported an open access model (Hopkins *et al.*, 2016) cited in (Tudur-Smith et al., 2015, p. 10).

In terms of time limits on access to study data, Research councils UK expect that data should be accessible for "*at least*" ten years after publication (UK Research and Innovation, 2018, p. 4), presumably post publication of the original research study, though there is a lack of clarity on whether this refers to the initial or final publication on the study in question. Tudur-Smith *et al* state that giving a "*reasonable*" indication to other researchers of when

data will be available post study is best practice, and quote 18 months as suggested by the Institute of Medicine as reasonable (Tudur-Smith *et al.*, 2014, p. 13).

3.9.3 Pros and cons of types:

The advantages and disadvantages of the two main approaches to data access for secondary researchers, open access and controlled access, are assessed briefly by some of the included guidance documents.

Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) do not recommend the use of open access sharing because individuals are at a greater risk of being identified (for example, if data are "*inappropriately merged with other data to facilitate re-identification*") and the difficulty in tracking publications and monitoring research arising from the data. If research projects arising from a data set are not made public or easily identified, researchers are not able to tell if research is being unnecessarily duplicated (Tudur-Smith *et al.*, 2015, p. 9).

Similarly, the Academy of Medical Sciences identified that open access bears the risk of identification but may also compromise the *"quality and appropriateness"* of secondary research conducted (The Academy of Medical Sciences, 2013, p. 2). A further limitation of the open access model considered was the "*reputational risk*" to the original data holder, for example if participants were not comfortable with commercial enterprises having access to "*data (originally) generated through government and charitable funding*" (The Academy of Medical Sciences, 2013, p. 2).

The only identified issue with controlled access models is that of resources. Requests need to be reviewed, and responded to, and may require further statistical support (The Academy of Medical Sciences, 2013). Data that is shared openly is perhaps considered less likely to attract queries, as it is further removed from the original research team. Data shared with controlled access will need access approvals to be assessed, may need data packs to be prepared, and have queries answered multiple times instead of once for open access.

However, both approaches require resources for data to be adequately prepared for sharing (i.e., cleaned, pseudonymised or anonymised as necessary, and documented). Ideally, both open and controlled access systems should seek to be standardised in terms of data management. For example, as pointed out by the Academy of Medical Sciences, "*data fields and attributes need to be subject to standardisation for meaningful data sharing to occur*" for both open and controlled access data (The Academy of Medical Sciences, 2013, p. 3).

The Academy of Medical Sciences report that participants spoke favourably of a third way, a system (for storing data) governed by an independent body with decisions on whether to share data made by an "*independent panel*" (The Academy of Medical Sciences, 2013, p. 3). The Academy of Medical Sciences explain that the details of the data requestor and the protocol of the proposed research should be studied by the panel (The Academy of Medical Sciences, 2013, p. 3). More detail on such independent panels is given below in Section 3.9.5 on Requests and Access Committees.

3.9.4 How to give access:

Six of the included guidance documents gave details on how researchers may give access to secondary researchers via prior preparation for sharing with policy and process documents in place. Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015, p. 12) provide a flow diagram of "*key data sharing activities*" throughout the duration of a study from trial inception to data storage that influence the success of data being shared at the end of it, such as the consent form, a data management plan, and database specifications (for the data pack).

The MRC also provide a flow diagram, where the emphasis is later in the sharing process, outlining the many stages of a formal request process for data sharing (Medical Research Council, 2011), highlighting that it is a cyclical process, with several stages of review prior to the production of a Data Sharing Agreement and eventual sharing. The first step of the flow diagram is a mechanism by which researchers can discover the study and its data. The National Institute for Health Research (NIHR) also require researchers receiving funding to publish a data management and access plan, as well as providing a *"data sharing statement"* alongside published research, detailing where and how research data may be accessed for secondary research (NIHR, 2019, p. 2).

As with NIHR (NIHR, 2019, p. 2) and MRC (Medical Research Council, 2011, p. 4), the principle that data are 'discoverable' is similarly advocated by the ESRC (and subsequently Research Councils UK) who suggest that all publications based on data resulting from an ESRC grant should specifically include information on where and how the data (and any supporting materials) can be accessed, ideally via a formal citation (ESRC, 2015(UK Research and Innovation, 2018). Any access restrictions should also be available to researchers at this time, having been set out in a data management and sharing plan *before* the original research begins (ESRC, 2015).

With regards to Data Management Plans, the Open Research Data Taskforce call for uniformity in funders' requirements for data management plans (whilst maintaining *"appropriate disciplinary differences")* and publishers' requirements for a data access statement in order to encourage consensus in the terms of secondary data use (Open Research Data Task Force, 2018). UKRI called for data management policies and plans that are in *"accordance with relevant standards and community best practice"* or *"national and international recommendations for best practice"*, leaving the researcher to identify the most appropriate standards if not provided by their funder (UK Research and Innovation, 2018, p. 4). Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) also refer to a data sharing policy that aligns with funders' or organisations' over-arching policies. Similarly, the NIHR simply state that data requests should be managed by following the policies for managing data requests of the organisation who has been contracted to carry out research by the NIHR, and that these policies must be *"transparent, robust, fair…"* (NIHR, 2019, p. 2).

Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) provide an outline of their suggested data request process, including use of application forms for data requests and review of requests against set eligibility criteria. The outcome of all requests "*with clear rationale for refusals*" should be made publicly available (Tudur-Smith *et al.*, 2015, p. 5). The estimated length of time taken to reach a decision on sharing should also be clear to applicants (Tudur-Smith *et al.*, 2015).

3.9.5 Requests and access committees:

As well as having data sharing plans and policies in place, for controlled access to data, research organisations are recommended to have access committees established to make decisions on granting access to researchers for secondary use. The guidance documents providing details on access committees were in agreement that the process of coming to a decision should be transparent regardless of exact access committee membership. There were several varying suggestions for composition of access committees.

Tudur-Smith *et al* suggest that decisions about data sharing in individual organisations may be made by *"data custodians"* such as members of the clinical trials unit, the Chief Investigator or statistician as described in the example case study of Keele CTU (Tudur-Smith *et al.*, 2015, p. 10). Tudur-Smith *et al*'s guidance points out that clinical trial units are often custodians of data, not owners; therefore, however impractical, the study sponsor should really be involved in data sharing decisions. Alternatively, recognising that decisions on

sharing may be biased when custodians are involved, requests can be referred to *"an independent review committee for a fully unbiased approach"* (Tudur-Smith *et al.*, 2015, p. 8). Contrary to Tudur-Smith, the ESRC state that data is usually 'owned' by the organisation conducting research on their behalf (ESRC, 2015) and the researchers conducting the research would therefore have to refer any decisions onto a committee to ensure an impartial sharing decision.

Other authorities also favour the use of data access committees (either a panel, or an individual) which are independent of the research project (The Academy of Medical Sciences, 2013), in lieu of decisions being made by data custodians (or original researchers). They indicate that such committees will need to review the protocol of the proposed research and the *"details of the requestor"* (The Academy of Medical Sciences, 2013, p. 3). The Academy of Medical Sciences refer to the ScottisH Informatics Programme (SHIP) (ScottisH Informatics Programme (SHIP), 2013) (now relocated to the Farr institute Scotland) which had developed "*a template to facilitate the concept of safe people, safe data and safe environment*". If the data requester could demonstrate all three components (safe people, data and environment) for the requested data, the access request was expedited (The Academy of Medical Sciences, 2013, p. 3).

Castell *et al* reported in their Ipsos Mori report that participants wanted (and therefore the report recommended) "*transparency*" and "*standardisation*" for access committees (Castell *et al.*, 2018, p. 32). They wanted access committees to "*prove their impartiality*" and "*demonstrate on what basis decisions are made*", and that committee composition should be consistent across the country, so that decisions are more likely to be standardised (Castell *et al.*, 2018, p. 33).

Castell *et al* went on to suggest that committee members should have no vested interests in research, or at least if they do, that these conflicts should be noted. It is not clear from Castell *et al* whether committee members should be independent of research in general, the primary research or of the proposed secondary research specifically, but some participants who took part in Castell *et al*'s research had suggested the *"need to mandate one-third lay membership as less than this was seen as tokenistic"* (Castell *et al.*, 2018, p. 32). Castell *et al*'s participants also had the, perhaps "*unrealistic*", desire that access committees should meet face to face to better facilitate *"debate and scrutiny"*, and in fact were still keen on this despite it being pointed out by the researchers that this might not always be practical

(Castell *et al.*, 2018, p. 32). This guidance document (Castell *et al.*, 2018), along with The Academy of Medical Sciences (The Academy of Medical Sciences, 2013) and Tudur-Smith *et al* (Tudur-Smith *et al.*, 2015) were the only documents that provided suggestions as to the composition of access committees.

The MRC policy and guidance goes further, proposing two types of access committees; Model 1 and Model 2, whereby Model 1 is useful if a study anticipate a large volume of sharing requests and Model 2 is more suited to occasional requests that are not likely to cause "*scientific, technical, ethical or legal issues*" (Medical Research Council, 2011, p. 27). Model 1 requests would be considered by a "*balanced*" access committee (one with a *"set of scientists with expertise of the purposes for which data are likely to be requested, supplemented ad hoc by further experts as required"*) within a given timescale and may call upon expert members to inform the Principal (*sic*) Investigator and their team (Medical Research Council, 2011, p. 27). Model 2 requests would be reviewed by the study team, with advice sought from a committee or single member at times of difficulty. The study team's decisions could be reviewed on an annual basis by a committee to ensure the decisions being made are sound (Medical Research Council, 2011). This approach encompasses data custodians and independent committees as practical and available.

For a data request to be processed, the reason for the data request needs to be stated with agreement within the access committee about what constitutes an "*acceptable*" reason for sharing (The Academy of Medical Sciences, 2013, p. 3). The Academy of Medical Sciences dinner discussion attendees identified a range of motivations for requesting data such as "*replication or verification of findings, seeking opportunities for collaboration, and new hypothesis generation*" (The Academy of Medical Sciences, 2013, p. 3). Prior to the involvement of an access committee, the MRC suggest that there should be a "*process in place to check both the authenticity of the researcher and their planned research*" (Medical Research Council, 2017, p. 22) or as, it is sometimes expressed, ensure sharing is with "*bona-fide research(ers)*" (Medical Research Council, 2011, p. 13; Corti *et al.*, 2014, p. 2).

The role of Research Ethics Committees in oversight of research was also briefly mentioned in the recommendations of Castell *et al*, where they are said to provide participants with reassurance that secondary research would be "*ethical*" with decisions on sharing independent and transparent, with emphasis placed upon the *"neutrality"* of committees (Castell *et al.*, 2018, p. 46).

3.9.6 Preparing data to share:

Some of the included guidance documents (Medical Research Council, 2011; Tudur-Smith *et al.*, 2015; The Academy of Medical Sciences, 2016; Open Research Data Task Force, 2018; UK Research and Innovation, 2018) made reference to the usability of the data actually shared, and how this is impacted by the accompanying metadata. Metadata refers to information accompanying the dataset that explains *"the origin, purpose, time reference, geographic location, creator, access conditions and terms of use of a data collection"* (Corti *et al.*, 2014). It was also acknowledged that the process of producing the data and metadata required specific skills and resources with associated costs.

The MRC and Tudur-Smith *et al* both point out that the staff preparing and working with the data at its home institution need to have the "*relevant knowledge and expertise*" to support the "*reasonable understanding and use of study datasets by new and external researchers"* (Medical Research Council, 2011, p. 16). This includes the ability to compile and provide useful metadata to accompany the data set (Medical Research Council, 2011) which can be achieved via an *"understanding of data management, basic statistics"* and the study itself (Tudur-Smith *et al.*, 2015, p. 6). The Open Research Data Taskforce also suggest the need to "*ensure that researchers have the necessary skills, along with helpful technical services and support from specialist staff*" when preparing research data for sharing (Open Research Data Task Force, 2018). Subsequently, *"quality control"* of prepared data can then be overseen by a *"further individual who is independent of the process"* (Tudur-Smith *et al.*, 2015, p. 6).

UKRI (UK Research and Innovation, 2018) and the Academy of Medical Sciences (The Academy of Medical Sciences, 2016) identify the monetary costs associated with production of data sets and metadata (staff time, hardware, software) and ensuring that data remains accessible via third party storage systems; they outline three "*cost recovery mechanisms*" for research grants (UK Research and Innovation, 2018, p. 11). The three potential elements in recouping costs associated with preparing and storing data outlined by Research Councils UK are *"directly incurred costs", "directly allocated costs",* and *"indirect costs"* which can be better achieved with *"effective use of data management plans"* (UK Research and Innovation, 2018, p. 11). Directly incurred costs are covered by study grants, and so any aspect of data management to facilitate sharing can reasonably be included in this. Directly allocated costs need to be justified and should only cover sharing activities that occur during the time of the original study. Indirect costs would cover data management activity that was

ongoing and related to all studies so could be used to cover infrastructure or administrative costs related to sharing (UK Research and Innovation, 2018). Tudur-Smith *et al*'s good practice principles guidance concurs that preparing for sharing can be resource intensive but warns that, although "*reasonable costs*" can be recovered, data sharing activity should not be profit driven (Tudur-Smith *et al.*, 2015, p. 15).

Once it has been established that the staff producing data sets are appropriately qualified and that costs are being recouped, the quality and utility of the dataset produced for sharing needs to be considered. The quality of the metadata produced can influence the "*re-usability*" of the data, and organisations should state: "*why, when, where and how?*" data was created and how it has been subsequently manipulated (Open Research Data Task Force, 2018). The Academy of Medical Sciences "*stressed*" that contextual knowledge of the study is "*critical*" and attempting to conduct secondary analysis without this knowledge could result in "*incorrect interpretations and misleading secondary analyses*" (The Academy of Medical Sciences, 2016, p. 2). To ensure "*understandability*", Research Councils UK also call for "*sufficient metadata*" and other relevant documentation to allow other researchers to interpret and re-use the research data (UK Research and Innovation, 2018, p. 4). Adequate metadata reduces the risk of *"unintentional misuse, misinterpretation or confusion"* (UK Research and Innovation, 2018, p. 5).

The Academy of Medical Sciences (The Academy of Medical Sciences, 2016, p. 2) publication is the only guidance document to refer also to data standardisation or *"universally recognised data standards"* to increase the utility of shared data, with reference made to the specific data standardisation proposed by the Comet initiative (Williamson *et al.*, 2017). Further details regarding an alternative data standardisation for research data can be found via the CDISC website https://www.cdisc.org/.

UKRI suggest that a delay in the release of data whilst "*sufficient preparation*" of comprehensive metadata takes place is acceptable, but that this should not be used as a reason for withholding access to data for secondary use (UK Research and Innovation, 2018, p. 8). To prevent undue delays to sharing, a research data management plan which allocates in advance "*sufficient resource*" for data preparation should mean that data can be shared within a reasonable time frame (UK Research and Innovation, 2018, p. 8). The guidance does not identify what would occur if the original researchers who are familiar with the dataset or the required resources are not available.

3.9.7 Data sharing agreements:

Where data sharing agreements were mentioned in the included guidance documents, the authors were unanimous that they were a useful way to facilitate transfer of data and conditions under which it can be re-used.

Once sharing has been approved and agreed, the MRC suggest that the Principal (*sic)* investigator should work together with a legal team to produce a Data Transfer or Sharing Agreement to be completed when sharing with other organisations (Medical Research Council, 2017). Tudur-Smith *et al* refer to this as a "*Data Use Agreement*" (Tudur-Smith *et al.*, 2015, p. 15). Research Councils UK specify that a data sharing agreement should be in place and signed by "*appropriate authorities*" (UK Research and Innovation, 2018, p. 6). Agreements should be in place before any data are released or any analysis performed (Medical Research Council, 2011; UK Research and Innovation, 2018).

Such agreements should set out *"what data are to be supplied, how they can be used and what will happen to these data when the research project complete"* (Medical Research Council, 2017, p. 22). For example, data may be required to be returned, destroyed, or access may be revoked (Tudur-Smith *et al.*, 2015). Data sharing agreements can also contain *"clauses"* that prevent the recipients from further sharing the data, making unapproved contact with the study participants or breaching confidentiality, e.g.: by attempting to identify individuals, presumably by linking the data with other data sets (Medical Research Council, 2011; Medical Research Council, 2017; UK Research and Innovation, 2018). These agreements may also detail any penalties for non-compliance with their terms (Tudur-Smith *et al.*, 2015).

Tudur-Smith *et al* also suggests that data use agreements can address plans for outputs from the study, for example ensuring that the original researchers are acknowledged (Tudur-Smith *et al.*, 2015). An example template for a data use agreement is provided by the Tudur-Smith *et al* in the good practice principles guide.

3.10 Guidance on type of sharing/trust?

The guidance documents were also searched for references to the type of sharing to be done, and trust of participants in research and researchers, including how this trust can be established and maintained. The Systematic Review (Chapter 2 and (Howe *et al.*, 2018)) identified that trust in researchers was an important aspect of consenting to share, and that

participants had preferences regarding the types of organisations that their data should be shared with. Communication with research participants was amongst the factors seen to foster trust in researchers.

3.10.1 Feedback to participants:

Five of the included guidance documents refer to provision of feedback to participants about how their data has been used. Although related primarily to personal, identifiable data (not anonymised and therefore unlikely to arise from clinical trials or health studies) the ICO state that *"You must ensure that individuals know what is happening to their data"* and should participants ask, *"you must also inform the individual about those organisations that you have shared their data with"* (Information Commissioners Office, 2020). The Academy of Medical Sciences consider feedback to participants regarding secondary analysis (where feasible) to be good practice (The Academy of Medical Sciences, 2013). It was, however, pointed out that not all participants wish to receive such feedback and therefore using the EU register for clinical trials to display information regarding *"who accessed the data for what purpose, and the results and any publications"* was suggested as an alternative method of providing feedback, although it is unlikely that participants would look there (The Academy of Medical Sciences, 2013, p. 3).

Castell *et al* identified that participants were in fact interested in what was done with their tissue (not data), for example how many times their sample had been used, but recognised that, due to time and resource implications, it may not be possible to provide individual feedback to each participant, or, presumably, if the option was given, to distinguish between those who had and had not agreed to secondary sharing (Castell *et al.*, 2018). However, there was still a desire by participants to have some general information and feedback about further research conducted with their data, and it was suggested by Castell *et al* that it could be made clearer and more transparent that further research was being carried out, perhaps with *"summaries on a website"* (Castell *et al.*, 2018, p. 25). Such decisions on how and when feedback (or the results of the research) will be given to participants should be made at the outset of a study (Medical Research Council, 2017).

The term '*feedback*' is also used by Castell *et al* and the MRC (Medical Research Council, 2017; Castell *et al.*, 2018) in the context of providing information to participants about their health, or *"unforeseen findings*" revealed in the original study. (Medical Research Council, 2017, p. 6). This is not relevant to this review and so is not covered here.

Similar to providing feedback on data sharing is the topic of reporting, identified by the Medical Research Council (Medical Research Council, 2011) where "*good reporting not only meets the requirements of researchers to be accountable to the funder for how they use public funds, but also enables their institutions and the funders to celebrate the success of studies and those who make secondary use of the data*" (Medical Research Council, 2011, p. 18).

3.10.2 Trust:

Trust in research and researchers was an important aspect of data sharing for participants as identified in the systematic review (Howe *et al.*, 2018), but was only mentioned as an important aspect of sharing by a small number of the included guidance documents. Lowrance (Lowrance, 2002, p. 66) refers to trust as being synonymous with the consent process and sets out the importance of a "*dialogue with the public*", to "*rebuild trust in the ways the NHS, healthcare professionals, and researchers, including academic and commercial researchers, use personal data, protect the data, and derive value from the data*". It is not clear from Lowrance's guidance which breach of this trust is being referred to by use of the term 'rebuild', although the guidance was published not very long after the Data Protection Act of 1998 came into force (UK Parliament, 1998).

Nevertheless, Lowrance devotes a section of the guidance to the ways in which researchers can "*nourish*" the public' or patients' trust, such as by understanding their concerns and preferences for data sharing, but also by coming to a consensus between organisations on a number of points such as the safeguarding of data and feedback on its use (Lowrance, 2002, p. 66).

The MRC refer much more briefly to trust, more specifically to the trust that is placed in researchers, reminding them that when secondary use of research data is granted, it is on the basis that they have been *entrusted* to "*deliver and responsibly communicate high quality research outcomes…respect the interests of cohort participants… ensure the integrity security and quality of information…*" and to "*properly acknowledge the original (*sic*) of the data and the significant contribution of various parties towards their creation*" (Medical Research Council, 2011, p. 21). The MRC's later guidance on using information about people in health research also refers to "trustworthiness", albeit briefly, with emphasis upon the management of data in a *"manner that demonstrates trustworthiness to maintain the confidence of participants and the population as a whole"* (Medical Research Council, 2017,

p. 5). Lowrance suggests that researchers enforce the commitment to safeguard data received as part of "*the deal*" of receiving it, and that, as they are usually using anonymised data and are therefore interested in "*cases*" not "*persons*", this provides reassurance of privacy and therefore trust (Lowrance, 2002, p. 66).

3.10.3 Bona fide researchers:

Several guidance documents (Medical Research Council, 2011; Tudur-Smith *et al.*, 2015; Medical Research Council, 2017) refer to the criteria which must be met by researchers before data can be shared with them for secondary research and agree that suitable recipients should be "*only qualified research groups*" (Tudur-Smith *et al.*, 2015, p. 13) or "*Bona-Fide*" researchers (Medical Research Council, 2011, p. 24).

The MRC (Medical Research Council, 2011) give a detailed description of their definition of 'bona fide' researchers, whose key characteristics are, briefly: "*an intention to generate new knowledge, using rigorous scientific methods"*, the *"professional expertise and experience"*, and the intention to publish the findings (Medical Research Council, 2011, p. 24). A bona fide research organisation would be one that has *"the capability to lead or participate in high quality, ethical research"*. Following this, Tudur-Smith *et al* suggest that a bona fide researcher can be "*evidence(d) via CVs and the involvement of a qualified statistician"* (Tudur-Smith *et al.*, 2015, p. 5).

In their later guidance, the MRC also refer to the expertise and qualifications which those handling information about participants (presumably the original researchers and those receiving data) must have prior to sharing; "*Principal Investigators* (sic) *must ensure that all those handling information about people (including students, visitors, collaborators etc.) have the relevant expertise in information security and local / study specific procedures*" (Medical Research Council, 2017, p. 25). Such staff should be appropriately trained in or have "*expertise*" in local information security policies and data responsibilities, including for control of and access to data, which should be defined before research starts (Medical Research Council, 2017, p. 25).

3.10.4 With whom data are shared:

Of the included guidance documents that mention with whom data might be shared, none preclude sharing of data with commercial organisations. UKRI encourage researchers to work in "*productive, equitable partnerships*", for example with charities and industry (UK Research and Innovation, 2018, p. 6). They go on to warn that, if researchers are working

with industry and the research results themselves are suitable for commercialisation, this should not "*preclude data-sharing and should not unduly delay it*", and a statement should still be made regarding where and how "*supporting data*" may be accessed (UK Research and Innovation, 2018, p. 6). Collaboration agreements should be drawn up between commercial and research organisations at the beginning of the study to set out who may have access to data in the future (UK Research and Innovation, 2018).

The MRC say comparatively little about commercial use of data (referring to it only in their Guidance on sharing of research data from population and patient studies), but do use the same wording as UKRI, also stating that they encourage researchers to *"work in productive equitable partnerships"* (Medical Research Council, 2017, p. 3). Both the MRC and UKRI use the statement that sharing with commercial organisations must *"conform to the same principles and practices as that required by the academic community"* (Medical Research Council, 2011, p. 3; UK Research and Innovation, 2018, p. 6). UKRI add that these standards *must also "conform to the requirements of relevant UK and EU legislation*" (UK Research and Innovation, 2018, p. 6).

The MRC add to this, stating that sharing with commercial organisations should be without exclusivity, for bona fide research and on a public interest basis (Medical Research Council, 2011, p. 3. p. 26).

Lowrance (Lowrance, 2002) agrees that research must be in the public interest, but that commercial research is just as likely to be in the public interest as that conducted by the public sector. Lowrance identified that the "*crossing of boundaries*" between public sector, commercial, and academic organisations and activities in healthcare and in research, can affect the way in which health data are handled, with particular reference to the development of "*hybrid databases under complicated custodianship*" (Lowrance, 2002, p. 7). When undertaking commercial research, UKRI remind us that "*all reasonable steps should be taken to ensure that research data are not held in any jurisdiction where the available legal safeguards provide lower levels of protection than are available in the UK*" (UK Research and Innovation, 2018, p. 17).

3.10.5 Research team recognition:

To recognise the "*intellectual contributions*" of researchers, but also to protect their *"intellectual property"* (UK Research and Innovation, 2018, pp. 7-8) most guidance made reference, however brief, to the acknowledgement and citation of the original research

team in publications arising from secondary research (Medical Research Council, 2011; The Academy of Medical Sciences, 2013; ESRC, 2015; Tudur-Smith *et al.*, 2015; HEFCE *et al.*, 2016; Cancer Research UK, 2017; UK Research and Innovation, 2018; NIHR, 2019). For UKRI this recognition dictates two of their six principles (principles 5&6) (UK Research and Innovation, 2018).

The ESRC suggest that this appropriate recognition can be achieved through the "*persistent information for citation*" accompanying all data that are deposited in a repository (ESRC, 2015, p. 2) whilst CRUK encourage the use of "*persistent identifiers*" such as Digital Object Identifiers (DOIs) and ORCID identifiers (Open Researcher and Contributor ID) (Cancer Research UK, 2017, p. 1). Tudur-Smith *et al* also suggest use of a DOI or co-authorship for original researchers in secondary research (Tudur-Smith *et al.*, 2015). Regarding acknowledgement, Cancer Research UK refer to the original researchers as "*data generators*" or as "*data sharers*" (Cancer Research UK, 2017, p. 1).

UKRI point out that researchers' intellectual property arising from research needs to be not only protected but *managed* in line with RCUK's "*Knowledge exchange principles*" (UK Research and Innovation, 2018, p.6). UKRI state that it is possible to delay data sharing for a "*reasonable period*" to implement measures that protect intellectual property such as filing of patent or licence applications (UK Research and Innovation, 2018, p.6). Research teams should also have the first opportunity to "*publish or otherwise exploit*" the results of their research (UK Research and Innovation, 2018, p.7). The Academy of Medical Sciences identified that funders and journals pushing for sharing as soon as possible after publication could have a negative impact on more junior members of a research team who often derive publications from *"subsequent analyses of clinical trial datasets generated by their team"* (The Academy of Medical Sciences, 2016, p.2). In these instances, a delay in sharing agreed with the journal or funder would benefit junior researchers.

UKRI do warn, however, that publication of research should not be unnecessarily delayed, as this could "*restrict the opportunities of others to use the same novel methodologies and/or datasets for other purposes*", and in fact, researchers could consider publishing the methodologies and datasets that are novel at the "*earliest opportunity*", even if they have not yet published their own findings (UK Research and Innovation, 2018, p. 8).

The Medical Research Council encourage "*productive*" and "*equitable*" partnerships where parties maintain their intellectual property, for example between researchers and medical charities or industry (Medical Research Council, 2011, p. 3).

3.11 Discussion

This literature review attempted to identify data sharing guidance or policy documents that could be utilized by researchers working in public health research (including interventions or longitudinal studies) or clinical trials.

Sixteen documents were identified as being relevant, and their guidance was summarised. The included guidance documents were a mixture of policies that must be adhered to and best practice suggestions, and a mix of those which were specific to data sharing, and those which were general data management policies which referred to data sharing in enough detail to warrant inclusion.

Most of the guidance documents cover the same ground concerning consent, storage of data or access to data, and some of the guidance documents also have overlapping or aligning policies (for example UKRI and the ESRC have overlapping policies as the ESRC are part of the UKRI, and similarly, the MRC co-authored Tudur-Smith *et al*), although they may be presented slightly differently or provide differing levels of depth on a particular subject matter.

The included documents were published between 2002 and 2020, and there seems to be little evolution in the guidance over these two decades. As demonstrated in Table 5, where one guidance document is lacking, another appears to fill this gap, but brings with it its own omissions in the guidance on sharing. For example, the MRC Data Sharing Policy (Medical Research Council, 2016) does not mention consent, but by referring to their later document, 'Using information about people in health research' (Medical Research Council, 2017) researchers can gain a fuller picture of that specific aspect, but no further information on access to data. Only two (Lowrance, 2002; HEFCE *et al.*, 2016) of the included documents comprehensively cover all four topic areas (consent, storage, access and type of sharing), one of which (Lowrance) is now quite dated, and both pre-date GDPR. For a complete overview of best practice on the key areas of importance to participants researchers may therefore have to consult multiple guidance documents.

What does seem to have emerged in more recently published guidance are calls for greater collaboration between bodies, and greater cross-reference to existing organisations and documents. This perhaps started in 2016 with the Concordat on Open Research Data (HEFCE *et al.*, 2016) and continued in 2018 with the update of the UKRI principles on data sharing (UK Research and Innovation, 2018) and the report of the Open Research Data Taskforce (Open Research Data Task Force, 2018). Indeed, the Open Research Data Taskforce explicitly aim to build upon the principles set out in the Concordat.

As also demonstrated in Table 5, it appears, although this may be coincidental, that more recent guidance becomes more specific and less broad, covering fewer of the four topic areas explored in this review than earlier documents did.

As expected, in most guidance documents the recommendations were skewed in favour of issues that concern researchers. For example, the ESRC guidance on research data management and sharing (ESRC, 2015, p. 2) presents six principles to which researchers whose projects are funded by them must adhere, and two of these (principles 5 and 6) concern researcher recognition or *"intellectual contributions"*. UKRI (UK Research and Innovation, 2018) refer several times to commercial constraints placed upon the sharing of data, which was not identified as a concern to participants in the systematic review (Chapter 2). Instead, participant concerns tend to focus upon the risk of having their data exposed or shared against their will with a commercial organisation whose principles do not align with their own (see Chapter 2 or (Howe *et al.*, 2018)).

The exceptions to the researcher-centric guidance were the Ipsos Mori *"public dialogue"* from Castell *et al* on behalf of the Human Tissue Authority (HTA) and Health Research Authority (HRA) (Castell *et al.*, 2018, p. 1), and The Academy of Medical Sciences (The Academy of Medical Sciences, 2013) publication, both of which included views of patient representatives. The inclusion of the HRA/HTA report and guidance in this review is questionable as it focuses to some extent on sharing of tissue samples and routine health records, but it was felt to be valuable in terms of the volume and participant-focussed aspect of guidance presented. The Academy of Medical Sciences is not strictly a guidance document; it is the notes from a dinner discussion that brought together experts in clinical trials, ethics, and data privacy, as well as patient representatives. It could be argued that, because of the inclusion of participant views, these two guidance documents are the most valid and useful to the focus of this thesis. Unfortunately, the Academy of Medical Sciences

have not gone on to publish any meaningful guidance on data sharing, though perhaps they considered this to be beyond their remit as an independent body.

Although most guidance focuses primarily on what we might term 'researcher issues', the key principles outlined do broadly align with participants' views (Howe *et al.*, 2018) that data sharing should be transparent, secure and with appropriate levels of anonymisation. A full discussion of the contrast between guidance and participants' preferences can be found in the discussion chapter of this thesis (Chapter 6. Discussion and recommendations for best practice).

A large degree of heterogeneity can be identified in the guidance documents featured in this review. There cannot be said to be any striking disagreements between organisations, but a researcher could still be forgiven for feeling confused in the face of the occasional varying approaches to and emphases on aspects of data sharing from different documents. For example, different recommendations are made regarding the most appropriate consent models from the UK Data Service (Corti *et al.*, 2014) and the MRC Hubs for Trials Methodology Research (Tudur-Smith *et al.*, 2015) with the former recommending that future uses of data be agreed prior to consent, meaning that consent is truly informed, and the latter not entirely precluding sharing data without any consent in place to do so.

The final report of the Open Research Data Taskforce acknowledges this lack of a standardised approach to sharing across organisations, identifying that the *"the current landscape is characterised by inconsistencies, gaps, overlaps, and lack of clarity – especially from the perspective of researchers – as to the roles and responsibilities of different organisations at local, national and international levels"* (Open Research Data Task Force, 2018, p. 30). Understandably, a researcher might simply follow the guidelines from their particular funder or institution, but this may not result in the best outcome for data sharing, data management or the participants whose data are collected as part of the study or in a consistent approach across studies.

Some funders or organisations provide guidance on selected aspects of data sharing but not others, for example, Cancer Research UK's guidance on sharing and preservation (Cancer Research UK, 2017) and The Medical Research Council's data sharing policy (Medical Research Council, 2016) make no reference whatsoever to the consent process. This makes it hard for researchers to gain a full picture of the best way to prepare for and implement

data sharing at the outset of a study, from a single source, or the source that may be most appropriate to them, for example their research funder.

3.12 Implications

The published guidance documents provide a useful starting point for researchers who want to identify best practice for data sharing. By utilizing this review, researchers can identify appropriate guidance on their area of concern, for example consent or anonymisation, if none is available from their specific funder. It also details the current situation in the UK at the time of writing (2021), with regards to data sharing guidance, and demonstrates that there is little overarching or definitive guidance that covers every area of sharing from conception of the study to study close, from which to draw on best practice. The Open Research Data Taskforce suggest that policy in the UK should be guided by the principles in the Concordat on Open Research Data (HEFCE *et al.*, 2016), but concedes that in practice it is not very widely referenced, and also fails to draw upon concepts such as the FAIR principles (Open Research Data Task Force, 2018, p. 42). The Concordat on Open Research Data is also now five years old and, like many 'guidance' documents, does not provide specific guidance on what researchers and research teams should do; rather it is more a set of principles by which research should ideally be conducted.

The Academy of Medical Sciences (The Academy of Medical Sciences, 2016, p. 3) called for a "*bottom-up*" collaborative approach to data sharing, modelled on current best practice. It is unclear, though, which organisation will rise to the challenge and combine all current best practice into one guidance document that can be utilized by researchers in the public health or clinical trials arena. Probably the most comprehensive document included in this review, that follows a study from inception to close and sharing of data, is Tudur-Smith *et al's* Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials (Tudur-Smith *et al.*, 2015), although this document is now six years old and, like many of the included guidance documents, seems to suffer from a slight hesitancy to be firm or absolute in its recommendations.

3.13 Limitations of the review

During the grey literature search it was difficult to identify documents that were wholly concerned with data sharing. It was easy to find guidance on consent for example, from the HTA/HRA (Castell *et al.*, 2018), but data sharing is sometimes only part of the story. Many documents were excluded from the review because they made only brief mentions of data

sharing and were more concerned with information governance or data management more generally. Some guidance that did not meet the inclusion criteria stated only the need for researchers to present a data management plan or data sharing statement, without elaborating on what these documents might cover. Considering the growing imperative from funders and publishers for researchers to share data and to include plans for data sharing as part of new studies, the lack of detailed and consistent guidance was surprising. Other documents were excluded because they were not from the UK or were collaborations between UK organisations and those of other countries, for example The OECD Principle and Guidelines for Access to Research Data from Public Funding (OECD, 2007). It is possible that useful guidance or recommendations that could be applied to a UK setting were missed in this way.

It was also difficult to identify a meaningful volume of up-to date-documentation. For example, Lowrance (Lowrance, 2002), dating back to 2002, is the oldest guidance document in the review, published presumably in response to a burgeoning data sharing culture and to the Data Protection Act of 1998. It could potentially be considered outdated following introduction of GDPR and in light of advances in technological capabilities for data sharing. This report refers the Data Protection Act itself, The Human Rights Act guidance from the ICO, the law and various bodies such as GMC and MRC.

Some documents, including Lowrance, mentioned above, did not make entirely clear the distinction between anonymised, pseudonymised and identifiable data, and at which type their guidance was aimed at or whether their guidance applied to all (e.g. (Lowrance, 2002; Medical Research Council, 2016; NIHR, 2019)). For example, the ESRC state that researchers should anonymise data or seek consent for sharing. The ICO guidance (Information Commissioners Office, 2020) deals almost exclusively with sharing of personal (identifiable) data, and the guidance is difficult to apply to the kind of data that results from trials, studies or interventions. However, it does apply to the data that must be held as part of running a study; for example, for contacting participants or sending out questionnaires, which should never be shared. It also applies to pseudonymised data. If the scoping review was conducted again it might be more appropriate to include documents which refer only to sharing of anonymised study data. It might also be easier for researchers to default to a stance of sharing only anonymised data.

Just two documents (Corti *et al.*, 2014; Medical Research Council, 2017) subtly distinguished between personally identifiable data for research study administrative purposes and research study data for sharing which may be identifiable or anonymised. Most guidance documents did not make this distinction, failing to mention identifiable data that must be used for running the original research study. Perhaps guidance documents should make clear their recommendations for administrative or personal data, which may need to be treated differently to that which forms the 'research' output.

Only three of the 16 documents refer to GDPR, and at the time of writing (August 2021), despite stating that it will be updated, the MRC guidance on using information about people in health research (Medical Research Council, 2017) has not been updated to reflect GDPR. The guidance in the grey literature review could therefore be considered to be less current than desirable, although one may imagine that the broad principles will remain the same following any updates.

The final report of the Open Research Data Taskforce (Open Research Data Task Force, 2018) sets out to build upon the principles set out in the Concordant on Open Research Data (HEFCE *et al.*, 2016), and therefore provides little in the way of additional useful guidance. It is also quite focussed on the funder and researcher, and on incentives and barriers to sharing, which were not the intended focus of the review.

A further limitation of the grey literature review was the decision to summarise the eligible documents within the framework of topics identified during the previously conducted systematic review. This could have led to important or relevant guidance that did not fall into the selected topic areas being omitted. It was assumed that guidance documents would, in general not make reference to participants' preferences or fears, but this may not be the case for all guidance. Castell et al (Castell *et al.*, 2018) for example, developed their guidance after consulting participants, so this addresses more directly their fears and concerns. A future review might benefit from attempting to identify areas where the guidance attempts to mitigate participants' fears and harms. When conducting this review, no guidance appeared to be directly appealing to participants preferences and concerns, but sometimes concerns were addressed as a by-product of recommended process such as the consistent emphasis on anonymisation and privacy. Future guidance could certainly benefit from incorporating practices that combat participants fears and harms as identified in the literature.

There is also a theoretical possibility that summarising of the included guidance documents could result in their messages being misrepresented. It is possible that quotes may, inadvertently, have been taken out of context, though every effort has been taken to avoid this. It is also possible, in principle at least, that a useful guidance document has been omitted either through deficiencies in the search strategy or the application of overly stringent inclusion criteria.

3.14 Chapter summary

This chapter described the process of conducting and the results of a scoping review of grey literature. To organise the guidance, it was categorised into the four main topic areas: consent, storage, access to data, and type of sharing.

Many of the included guidance documents refer to the planning stages of studies when researchers should begin preparations for sharing, (e.g. (Tudur-Smith *et al.*, 2015; Medical Research Council, 2017)). Planning to share at the outset of a study means that participants and researchers are prepared, and processes are in place, but, in reality, as in the guidance, research data sharing is still often an afterthought. Coupled with this is the fact that few guidance documents provide a comprehensive guide to all aspects of sharing throughout the study lifecycle such as consenting, storing, giving access, and providing feedback to participants, instead choosing to focus on specific aspects of sharing whilst ignoring others altogether. There is a need for a guidance document that details the processes and procedures required for sharing from study outset to study completion and beyond, guidance that goes beyond suggestion, and gives detailed descriptions of what researchers should do; a sort of data sharing manual.

Crucially, there is not a great deal of detail in the included guidance on aspects of sharing that participants are interested in; concerns such as having data shared with organisations whose principles do not align with their own, or which may try and engage them in unsolicited contact, or on desire for feedback and the opportunity to learn how their shared data has been used. There is indeed very little reference to research participants at all. The guidance does however broadly align with participants' preferences as identified in Chapter 2, recommending for example that consent should be sought as far as possible and that privacy should be protected. Privacy was a key participant concern, although it is not clear whether the guidance places emphasis on this because of participant concerns, or because this is a legal requirement.

To encompass participant concerns as well as the more practical researcher queries, future guidance documents should consider ways in which researchers can increase transparency at all stages of research, from a detailed and fully informed consent process, through to keeping participants aware of who their data is shared with, giving them the option to decline if consent allows, and providing feedback on study outcomes. Further work is also required in the research community to ensure that data sharing guidance documents align with one another, such that, no matter which funder or journal requires data sharing, it is carried out in the same way each time, whilst still allowing for individual circumstances where sharing may not be possible.

The following chapter details the development of a questionnaire survey which was informed by a scoping focus group.

Chapter 4 Questionnaire development- scoping focus group and questionnaire design

4.1 Introduction

In this chapter I detail the iterative process of questionnaire development including the scoping focus group (and results thereof) and cognitive interviewing undertaken as an essential first step in the production of a self-completion survey of attitudes towards data sharing. The chapter ends with a description of piloting of the questionnaire and a final draft ready for distribution.

4.2 Scoping Focus Group Methods

A scoping focus group (SFG) was conducted in October 2017 with two main purposes:

1) To determine whether the attitudes and preferences of research participants and members of the public regarding research data sharing, as identified in the systematic review of international literature (Chapter 2), were shared by UK participants; and
2) To contribute to the identification of topics and items for inclusion in the survey of attitudes to research data sharing.

A focus group was identified as being the most appropriate method for this scoping exercise due to its recognized function as an *"exploratory"* tool for *"hypothesis generation prior to developing a more structured questionnaire or interview, but also as a tool to identify people's views and understandings"* (Wilkinson, 1998, p. 184). By conducting a focus group, as opposed to a series of one-to-one interviews, it was possible to obtain *"several perspectives about the same topic"* in a short period of time (Gibbs, 1997, p. 1). It was also thought that the discussion or conversational nature (or the interaction between group members) of the focus groups would allow participants to explore their understanding and views of data sharing with others (Gibbs, 1997, p. 3). They would be able to use *"their own vocabulary…pursuing their own priorities"* (Kitzinger, 1995, p. 299) while simultaneously engaging in group interaction to *"produce data and insights that would be less accessible without the interaction found in a group"* (Morgan D. L, 1997) through *"collective sense making"* (Wilkinson, 1998, p. 186).

Although scoping groups are a useful tool for ascertaining views of participants, they are not without challenges. The researcher needs to be mindful of both facilitator and respondent bias and take steps to reduce this where possible. It is also hard to ensure the generalisability of results as samples are often small and unrepresentative (Gibbs, 1997; Wilkinson, 1998). In this particular instance, however, the focus group data collection was a means to an end, rather than an end in itself, and therefore generalisability was not its aim. Finally, although focus groups may seem *"naturalistic"*, the data have actually been gathered in a specific context (Green & Thorogood, 2014, p. 127).

4.3 Sample

For the scoping focus group, participants were recruited from VOICE (formerly VOICE North) (VOICE, 2017), an organisation based in the north-east of England which allows members of the public to actively contribute to research activity. Prior contact had already been made with the organisation to confirm that this was a possibility, and I provided a lay summary of the research for potential participants to be used in recruitment. An advertisement for recruitment, giving details of the study was placed by the organisation on their website and in their weekly newsletter, following ethical approval of the study by Newcastle University. Members of VOICE who expressed an interest in taking part were given the participant information sheet and consent form by a contact at VOICE.

Individuals who had actively sought to contribute to research, even if they had not actually participated in a research study, were considered suitable informants for this data collection exercise, as it was felt that they might have some insight into, or be more easily able to respond to, the issues around research data sharing. My systematic review on the topic of data sharing (Howe *et al.*, 2018) as well as research identified during the systematic review (Jao *et al.*, 2015b) had demonstrated that participants who were completely naïve about data sharing needed extra help to understand the concept before they could give their opinion.

Eligible participants were individuals aged over 18 years of age, regardless of whether they had previously taken part in a trial or public health research. The recruitment advert stated that direct or indirect experience of research, trials or studies would be an advantage but was in no way a requirement.

All participants who expressed interest in taking part (n=12) were invited to do so with the exception, due to the potential cost of reimbursing travel, of one who lived outside of the North East. One participant who wanted to contribute by email or telephone did not respond once the participant information sheet and consent form had been emailed to them, perhaps because the PIS and consent indicated that participation would be via a focus group.

This left ten participants who had agreed to attend the focus group, seven of whom attended on the day. Participants who attended were all of a similar age (59-75) and comprised five males and two females. All participants were white. Participants were not issued with name badges, to preserve anonymity, although some of the participants had met each other at previous VOICE activities; and introductions were made prior to sound recording. After reading McLafferty (McLafferty, 2004, p. 193), I anticipated that if participants were already familiar with each other, conversation might flow more easily as there would be a *"positive group dynamic"*. All could be said to have taken part in research previously (given their membership of VOICE), but only a few participants disclosed during discussion whether or not they had taken part in a research study previously and did not give specifics about the type of research this was. In order to preserve confidentiality, I did not ask for details.

Too few interested participants contacted VOICE within the given time frame for me to be able to screen and select participants based upon age or gender (although there was an almost even gender split in the 10 participants who intended to take part). It was also pointed out through discussions with the engagement team that the VOICE population tended to be older or retired and so a more representative sample age-wise might not have been possible.

4.4 Location

The focus group took place in the medical school at Newcastle University, a venue that is known to participants of VOICE. A quiet room was booked to ensure that there were likely to be no interruptions and that privacy was maintained. I was assisted on the day of the focus group by a colleague who guided participants from the medical school reception to the room where the group was taking place. Participants were provided with light refreshments and offered forms to claim a reimbursement of their travel expenses.

4.5 Consent, data protection and ethical issues

Individual written informed consent was obtained from participants prior to commencing the focus group. VOICE distributed Participant Information sheets and consent forms to participants who had confirmed their wish to take part in a focus group prior to them attending the focus group/interview, thereby giving them time to consider the information. Duplicate participant information sheets were also distributed to each individual on the day. Consent forms and participant information sheets were subjected to a readability test online (Scott, 2017) prior to submission for ethical approval; they were found to be suitable for a reading age of 13-15 years of age.

At the focus group, participants were given a brief description of the study and its aims. Participants were informed that they had the right to withdraw from the study at any point. They were made aware that the group discussion was to be recorded, that the transcription of the interview would be anonymised, and the recording erased after transcription, and that they would not be identifiable in any published work.

Participants were also reminded that they could obtain a summary of results of work to date via VOICE. Research team contact details were also provided to participants and the participants were informed that they could contact us (myself and my main supervisor) to ask questions about the study at any time.

Completed consent forms and any physical data resulting from the focus groups and interviews were stored in a lockable filing cabinet at Newcastle University. Electronic data was stored on a password protected PC within Newcastle University.

The participants were asked on the consent form whether they were comfortable with the data collected through the focus group potentially being shared in the future with other researchers. They were reassured that, prior to sharing, the data would be thoroughly anonymised, as its conversational nature might result in inadvertent identification of participants.

4.6 Materials and Method of Enquiry

Guidance on conduct of focus groups or group interviews was sought prior to undertaking the scoping focus group, for example: (Kitzinger, 1995; Kreuger and Casey, 2015). Data collection for the scoping focus group was carried out using a semi-structured approach with a schedule of open-ended questions to prompt discussion with participants in a

conversational style. To facilitate this, a focus group topic guide was produced; this was informed by the systematic review (Howe *et al.*, 2018) and by any topic guides available as supplementary files in studies included in that review (e.g. (Asai *et al.*, 2002; Jao *et al.*, 2015b; Manhas *et al.*, 2016)). The use of open-ended questions was designed to get participants to expand upon their answers, starting with general topics and becoming more focused as the discussion progressed (Green & Thorogood, 2014). The focus group topic guide had previously been submitted to and approved by Newcastle University ethics (see Appendix D Scoping Focus Group Topic Guide).

The open-ended questions in the topic guide were designed to be delivered carefully in an attempt not to influence the participants' answers; any prompts to participants were as neutral as possible in a conversational setting. I attempted to let participants chat amongst themselves, a sort of *"structured eavesdropping"* (Kitzinger, 1995, p. 301), using neutral prompts to get participants to expand upon any points raised, whilst avoiding leading them in any one direction. It is also the facilitator's job to move participants on to the next topic when conversation drifts or they seem to have reached a *"minor conclusion"* (Gibbs, 1997, p. 5). Allowing participants to talk amongst themselves when exploring the issue of data sharing reduced any potential influence that I might exert on participant responses, as noted by Wilkinson (Wilkinson, 1998, p. 190), although on occasion it was necessary to use more focused questions with a little explanation when the participants went off track or needed help to understand what was being asked.

Participants were also provided with reference materials in the form of example scenarios to explain the concept of data sharing, to help them form and give their opinions. They were given the opportunity to discuss scenarios that were most important to themselves ensuring that the agenda was not purely my own or too closely tied to the topic guide, ignoring participants' priorities. Presenting participants with scenarios not only helps with understanding the topic but encourages participants to focus on each other and the topic to be discussed rather than looking at the facilitator for cues (Kitzinger, 1995, p. 301). The scenarios presented to participants were also based on simplified versions of those provided by other studies into data sharing (Jao *et al.*, 2015a; Jao *et al.*, 2015b; Merson *et al.*, 2015; Manhas *et al.*, 2016).

4.7 Reflections on the focus group

Prior to conducting the focus group, I had attended an MSc module on qualitative techniques run by the Institute of Health and Society at Newcastle University (now Population Health Sciences), which included best practice for interview and focus group techniques. However, I had never conducted my own focus group and was worried that my inexperience may be apparent to the participants and ultimately influence the data collected.

It is possible that a researcher who was better practiced at facilitating focus groups, for example knowing when to provide prompts or questions at the right time and when to stay silent could have elicited a 'better' response or more in-depth discussion around certain points and less repetition on other points. Sim (Sim, 1998, p. 347) describes the role of the facilitator as *"pivotal*" to the nature and quality of the data collected, and Agar and MacDonald (1995) warn that moderator control can have an *"important effect on the quality of group discussion"*, with too much control preventing discussion and too little control resulting in a topic not being discussed at all (Agar, M. and MacDonald, J., 1995) cited in (McLafferty, 2004, p. 192). Sim refers to this is striking the right balance between an *"active or passive role"* (Sim, 1998, p. 347). I continued to refer to the topic guide throughout the focus group, but it was sometimes necessary to amend the types or order of questions as the discussion progressed in reaction to the participants discussion. Towards the end of the focus group, participants were asked to consider data repositories, and it was at this point that the discussion seemed to continue beyond the stage that it might have naturally concluded without my intervention. It is possible that a more experienced facilitator may have been able to elicit more information about repositories or storage.

I did however receive positive feedback in writing from one participant, and other participants commented that they had found the session interesting. From my perspective the experience overall was rewarding, in terms of volume of data achieved, the interaction with participants and the experience gained.

4.8 Data processing

The scoping focus group was digitally recorded and transcribed verbatim by me. In line with GDPR (Information Commissioners Office, 2019b) all identifying information was anonymised. When transcribing, the identity of each individual was replaced with a number

between 1 and 7. Any references by one participant to another in the focus group transcription were replaced with the corresponding number. During transcription it was relatively easy to determine which of the seven participants was speaking as I had grown familiar with their voices. There was no second facilitator taking notes but given the small number of attendees this was fortunately not necessary.

Care was taken to follow guidance on transcription to minimize bias that could be introduced by misrepresenting quotes or discussion (Transcribe.com, 2015). The focus group was described verbatim, including *"non-verbal utterances"* (Transcribe.com, 2015) such as 'umms' and laughter, which helped me to be able to record whether something was said in jest or with sarcasm. I relied upon the fact that I had spent time with the participants during the focus group to be able to determine when they were being sarcastic for example, and by listening to the change in voice tone. This assisted with coding, as it was important to note *how* things were said, as well as the content. Any mispronounced words, hesitation, interruptions, or drawn-out words were also transcribed as heard, to avoid changing the nature of the discussion. Time stamps were included throughout the transcription, at first to increase the ease of transcription, but once the original recording was deleted, the time stamps were kept in the document to provide an accurate record of the pace of discussion.

One element of focus group analysis that was not incorporated into the transcription was recognition of the impact of *"the group dynamic"* (Kitzinger, 1995, p. 301), in that little notice was taken of the interaction between participants. There are several reasons for this. The first is that the purpose of the focus group was purely as a scoping exercise and to identify themes. It was not important at this stage whether participants influenced each other's attitudes or who was responding to whom. Secondly, when listening back to the recording there did not seem to be a great deal of interaction between participants, instead participants seemed to consider points raised and then give their opinion, in turn, at an appropriate point. This might have been because of the presence of one or two dominant personalities in the group causing dissenting views to be *"artificially suppressed"* (Sim, 1998, p. 348). This also might have been because all participants were genuinely in agreement with each other. Finally, it is important to remember that focus groups are discussion *"occurring in a specific, controlled setting"* as opposed to a natural conversation (Smithson, 2000, p. 105) and should be analysed as such.

4.9 Coding and analysis

The transcript was imported into NVivo V10 (NVivo, 2021), to facilitate a brief thematic analysis of prominent or recurrent themes in the data. Coding was then conducted; this simply means selecting words or phrases of interest that become free codes (n=227) known as 'nodes' in NVivo. These words or phrases were those which demonstrated opinions on data sharing, or words or concepts which were repeated often, were novel, or struck a chord with me for an unknown reason. For me this coding process was instinctive.

Each node was grouped with like nodes to form grouped nodes (n=35), for example, comments about being identified in data would be grouped together under the heading 'identification' (see Figure 6 below). In the Nvivo software this literally means cutting and pasting nodes into groups with similar nodes/quotes and re-arranging or renaming groups of nodes as necessary/until satisfied.

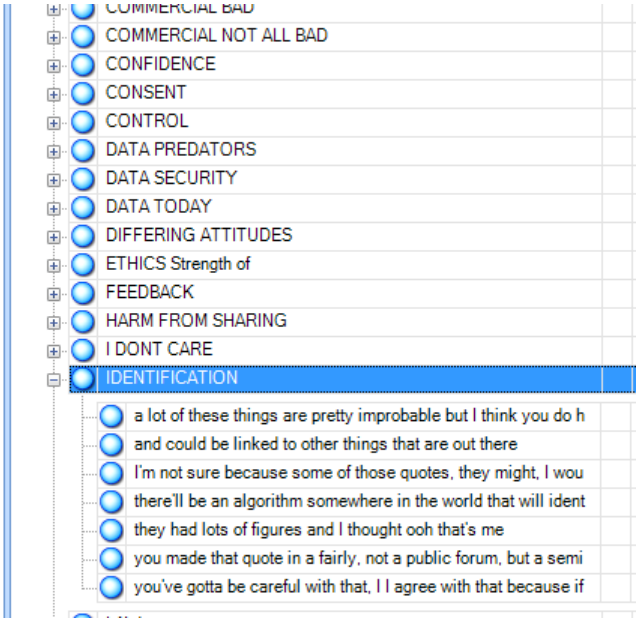


Figure 6: NVivo screenshot showing a selection of grouped nodes with one group expanded to show initial nodes.

The groups of nodes were then further amalgamated in Nvivo into 11 groups (Figure 7), for example, the group of nodes referring to 'identification' was grouped with the group of nodes 'unwanted contact' as these are both considered to be negative consequences of data sharing. These groups were joined by the general 'harm from sharing' group and overall given a name chosen by me. I decided that 'harm from sharing' was a good catch all, and it also appears in previous literature.

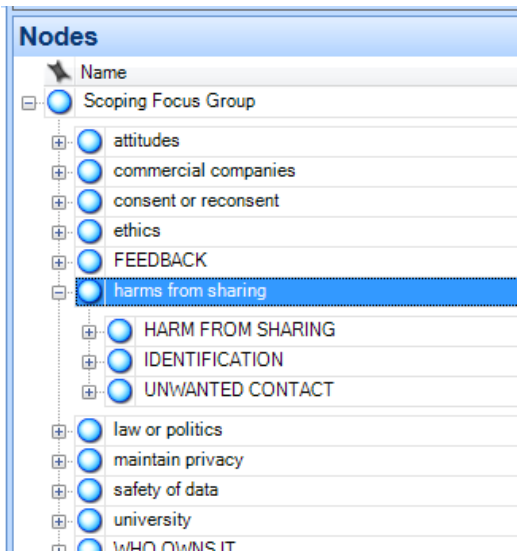


Figure 7: Grouped groups in NVivo

Groups were then combined until they were broad enough to be considered themes. For example, the final theme 'How Sharing Affects me' encompasses both the positive and negative aspects of sharing. The title of the theme was determined *after* the groups had been arranged. Four main themes were decided upon and they were then given titles that were reflective of their content. These themes are detailed below in the results section of this chapter (see results section 4.9), while Figure 8, below, shows a single expanded theme, and how it encompasses individual nodes.

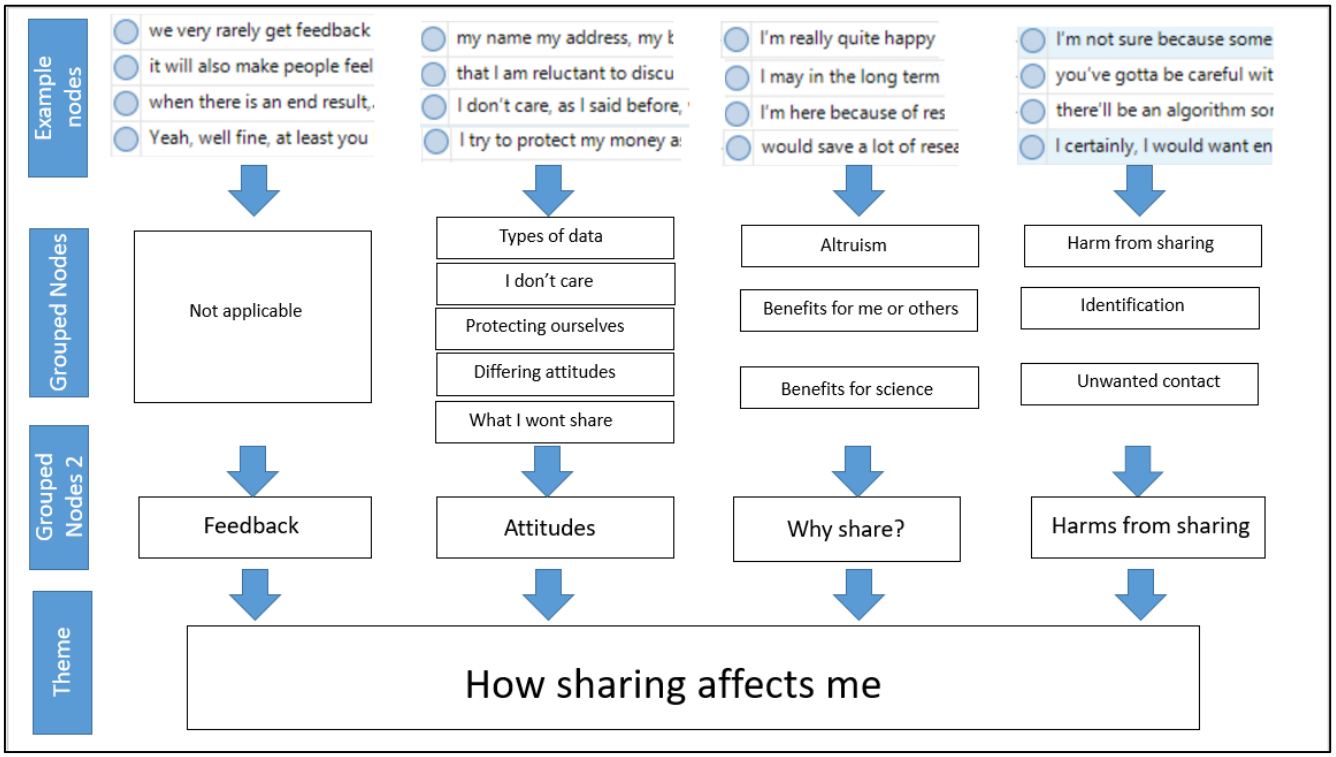


Figure 8: Expanded theme to show 'nodes' that informed it.

Themes are reported thematically as per Thomas and Harden (2008) (Thomas and Harden, 2008), but no attempt was made to provide reasons for the existence of the themes. Data collection and analysis was an iterative process that allowed for codes to be grouped and re-arranged until I was satisfied.

I decided not to prepare an a priori framework (Thomas and Harden, 2008, p. 4) to code the transcript; rather the coding was inductive as far as it could be, given that the prior experience of coding systematic reviews exploring participants' attitudes towards data sharing, (and indeed the experience of sitting in the focus group itself), meant that I may have held pre-conceived ideas of what themes may emerge! This may have influenced the words or phrases that were chosen to become codes and subsequently the themes that emerged from the codes. The themes identified in the systematic review also knowingly influenced the type of questions asked of participants in the focus group. However, as long as this is acknowledged, along with the fact that the focus group was purely a *scoping* exercise to explore whether participants agree or not with the views of participants detailed in published data, this was deemed not to detract from the usefulness of this component of the research.

Care was taken during coding to distinguish between jokes and general discussion, to ensure that codes (nodes) were not taken out of context, referring back to the original transcripts where necessary. It was hoped that, by coding individual words or phrases from the transcripts, no more importance was given to one concept over another, regardless of whether opinions were those of more or less-vocal individuals, minority views or group consensus, and that codes were not deliberately misinterpreted by me, the researcher, to fit into expected themes or theory.

The themes identified during the focus group were used to inform questionnaire development, where it was intended that each identified theme would be covered in the questionnaires. It was further intended that the language and expression used by participants in the scoping focus group would also be reflected in the wording of questionnaire questions, ensuring language that lay participants would relate to. For this reason, using a focus group to "*provide access to participants' own language concepts and concerns"* (Wilkinson, 1998, p. 181) was particularly appropriate.

4.10 Results of thematic analysis of scoping focus group

The thematic analysis of the scoping focus group identified four main themes;

1) The Nature of Data Today, encompassing data security, linkage, legislation and sharing with commercial bodies;
2) How to Maintain Privacy Ethically, including anonymisation, the role of ethics committees and the consent process;
3) Different Users, Different Trust, referring to the differing levels of trust in different institutions; and
4) How Sharing Affects Me, exploring how data sharing has the potential to cause harm, but also to benefit the individual and society.

A feedback summary of the themes identified was sent to the VOICE Members who took part, and a summary of their participation was included on the VOICE website as a blog post available at: https://www.VOICE-global.org/latest/2018/may/VOICE-members-views-on-data-sharing/

4.11 The Nature of Data Today

Participants made frequent references to the way in which shared data might be traced back to individuals, linked with other data and used for unsolicited purposes by commercial bodies. There were questions asked about the security of any data shared, given the way in which technology allows information to be easily shared or found online. One participant reported that *"confidence"* levels regarding safety of data varied by institution, and this notion of confidence was then taken up and repeated by other participants.

4.11.1 Commercial organisations

As part of the discussion, many references were made to *"commercial organisations"*, (prompted initially by the scenarios given to participants), with whom research data could be shared. The trust exhibited in universities to protect data was not extended to commercial bodies such as *"pharma companies"*. Concerns were voiced about organisations being *"unscrupulous"* and potentially *"manipulating data"* to suit their own needs or using it to *"sell insurance".* Commercial organisations were also seen as being more likely to be *"taken over"* by other companies, meaning that any assurances in place regarding data privacy may be neglected:

your information could be spread around the world.

Some participants were clear that they would not be happy for a commercial organization to use their data without consent, and one mentioned an instance where they had inadvertently taken part in research that was for commercial benefit.

I was quite upset about that...I wasn't happy at all.

However, not all connections with commercial companies were viewed negatively. There was recognition that sometimes partnership between universities and commercial bodies could advance research

I mean universities now have to have commercial partners to survive to get their research on the market.

This benefits patients in the long run, and one participant gave the following example of the blurred line between commercial and non-commercial (academic) research:

a drug company that wants to use that data to produce drugs that would help people in the long term.

4.11.2 Law/politics

Several references were made to the law surrounding data protection and how these laws are potentially liable to be amended for political reasons:

…so I make up laws which say its ok we'll sell this information to this pharmaceutical company, hey it's in the public interest guys and all the rest of it, but really it's for my own personal gain.

It was mentioned that agreeing to sharing under one law may mean that in future, data could be shared in a way not initially agreed:

if our information is put on file, and it's used again in 5 years or ten years' time, the law may have changed again.

4.11.3 Data security/predators

From some participants, there was an overall lack of trust in the security of data, data online and of data controllers or users, with references made to *"data predators",* data as a weapon and *"data giants"* such as Facebook, Twitter, and Google. Participants made references to data security, citing the strength of IT systems, NHS data breaches, ransomware, cloud storage, insecure web browsers and anti-virus software as potential areas of concern

> *you know, six steps down the line somebody's IT system may not even have an anti-virus.*

The idea that research data could potentially be linked with other data freely available online was brought up by a couple of participants, with references made to social networking services such as Facebook where the younger generation in particular were seen to share information freely:

> *"I think the concept of privacy has virtually disappeared from the younger generation".*

It was perceived by one participant that even in research papers (with anonymised and/or grouped data) the individuals involved could potentially be identified by looking for published reports online if someone had the time and inclination to search for it or knew which individual they were looking for:

> *so, the research is there, the people who've taken part in it, it's all there for money or for nothing if you're a good detective on the internet.*

It was suggested that nothing is really 'private' anymore, both in terms of doubts over data security and due to the predatory nature or some groups or individuals. Participants suggested that those agreeing to take part in university-led research should consider the wider context at the time:

> *do I want to pass it to a university, and look at the world out there and what's happening to data…*

All of the above discussion was tied to the notion of "confidence" in the storage and usage of shared data, which could depend upon the organisation(s) involved:

> *the level of confidence in the holding of data these days is being considerably reduced.*

Confidence would *"wane"* the less the individual knew about where the data was going or what it would be used for, and there was a notion that participants should not just consider that data is being given to the intended sharer, but to the world:

> *confidence in when you give data to the world that the world looks after it.*

4.12 How to Maintain Privacy Ethically

Participants stated that they wanted their data to be sufficiently protected if it were to be shared. They were interested in the process by which decisions to share would be made, enquiring for example about ethics committees and the type of data governance in place at universities. Participants felt that they had the right to privacy, and therefore those collecting or sharing data had responsibilities to maintain that privacy.

Participants discussed the consent process (when prompted) and the idea of re-consent if data were to be shared. The concept was explained to them, giving the example of a university researcher contacting them to check if they would agree to share their data with another researcher or organisation.

4.12.1 Consent

Some participants mentioned that the consent form (and anything it may say about sharing) was of little importance if they had already decided to take part in a research study:

> *Frankly..., I wouldn't read it properly.*

Participants advocated being as clear as possible on the consent form about with whom data would be shared so that they were informed at the beginning of the study:

> *The basics at the beginning...surely if you want to recruit us, you put on your paper... and that gives us an idea of where the data's going to go.*

Participants suggested that researchers could use caveats for sharing on the consent form, should they decide in future to share with an organisation or project not pre-approved by the participant:

> *you can put on your consent form we will do our best...but nothing's perfect….*

Conversely, participants were also aware that the nature of data sharing can be too unknown to give details at time of consent but that this might deter some participants from taking part:

> *It is difficult because…if you had a very very broad consent form, some people, maybe me included would be less likely to consent to it. Because there's less certainty.*

In conclusion, one participant pointed out that the way the consent form is phrased is likely to be key to encouraging sharing without damaging participation in the original study.

4.12.2 Re-consent

When it came to re-consent, participants were pragmatic, and discussed both the pros and cons of such a process. Participants indicated that it would be a *"nice idea"* to be approached for re-consent and that it *"sounds sensible"*, but on the whole the group agreed that re-consent was *"a bit of a can of worms" "unwieldly"* or *"out of hand".* They questioned how researchers would find or get in touch with participants if they had moved home or changed email address and that those taking part in research *"… expect all kind of things like this"* (data to be shared). Separating participants who had and had not consented to sharing was seen to *"complicate matters"*. Participants understood that to use data only of participants who had consented to sharing would be to introduce bias into any subsequent research:

> *...you've sort of committed yourself to a certain group of people haven't you really.*

One participant suggested that as a sort of compromise:

> *you could say well you preferred it not to be but unfortunately, we had to do this (share).*

4.12.3 Ethics

Several participants referred to, and were interested in, the way in which the ethics process or ethics committees may oversee the potential sharing of study data, predominantly using the example of universities sharing research data:

> *If you'd put on this bit of paper, this has been approved by the university ethics committee, what does that mean?*

They wanted to know that *"those ethical hurdles are adequate"* and that they'd be happier,

> *…if at the university level there was another kind of ethical…barrier before you could pass your information on.*

One participant suggested that having access to details of an organisation's ethical standards (however unfeasible this might be in practice) would help them to decide whether or not to agree to share data. The adequacy of the ethical *"hurdles"* was questioned when it came to data safety, with one participant asking:

what sort of ethical constraints are put on the commercial organisations that partner with the university?

There were brief references made to the ways in which we could learn about good data sharing practice and process from other organisations or countries.

4.12.4 Responsibility to ensure privacy

There was emphasis placed on the university's responsibility to keep data confidential and make sensible sharing decisions that would not threaten the participant's right to remain anonymous:

well, my feeling is...this should be the university's responsibility.

Maintaining privacy or not *"giving away your privacy"* was key to some participants, through anonymisation or sufficient controls regarding with whom data was shared. It was pointed out by one participant that:

we have the right to say to you, I don't want that data to be used.

Despite participants claiming on the whole to be open to the idea of sharing their research data with secondary researchers, when asked they were able to suggest potential concerns about privacy and maintenance of anonymity that, although perhaps not an issue for them personally, may be of concern to other participants in other settings. Perhaps participants were not concerned initially concerned about being identified because of the understanding that all data would be anonymised before sharing:

so, you couldn't identify an individual, it's just the results of that research?

Those with whom data might be shared should also maintain this anonymity:

as long as it can't be traced back to you, I don't see the problem frankly.

4.13 Different Users, Different Trust

Participants exhibited a high degree of trust in a university but were less confident about taking data outside of a university. Participants discussed at length the types of users or organisations that they would and would not trust with their data, and one participant asked:

who owns the data after you've done the research?

4.13.1 Trust within and outside the university- who will it be shared with?

Although the majority of participants initially stated that they did not mind what happened to their data once they had taken part in a study, prompts during discussion did encourage them to consider instances where they would prefer their data not to be shared, and these largely concerned the *type* of organisation to which data would be given.

Participants exhibited a high degree of trust in a university to share their data responsibly and do the right thing *"I would have a lot of faith in them".* Sharing the data with researchers or other universities was acceptable if:

> *the data's going to be shared among other researchers, then that's fine*

However, once the data was shared outside of the university, there was a greater degree of uncertainty both about maintenance of privacy and sharing for purposes previously approved by the participants. This may trigger a desire for re-consent

> *It would be useful to know which projects that information was being used for.*

4.14 How Sharing Affects Me

Participants discussed the reasons for sharing data and the benefits to both the scientific and research communities and to patients who should benefit in the long term. The participants in the scoping group had varying attitudes towards sharing but all were able to identify potential harms that could befall them or other participants as a result of irresponsible data sharing.

4.14.1 Potential harms from sharing

Potential harms from sharing data largely concerned unwanted contact from organisations or inadvertent identification through not protecting data thoroughly

> *"I think sometimes you can say sorry, but the damage has already been done".*

Despite most participants being happy for researchers to share data, they were not agreeable with being identified and then contacted with advertisements or as one participant put it:

> *maybe a faith healer or something trying to come to my doorstep…*

Identification was also discussed with participants agreeing that it was *probably* unlikely but nonetheless possible that an individual could be identified in anonymised data, particularly if they had a rare disease or set of circumstances:

> *there'll be an algorithm somewhere in the world that will identify you*

Participants were asked to consider whether they would mind text from discussion or quotes being shared and one participant stated that they would exercise caution because:

> *"…some of those quotes, they might, I wouldn't be identified as a person, but they might reflect upon me as a personality"*

Another participant considered that quotations were fair game for sharing because:

> *you made that quote in a…semi-public forum, among strangers, therefore that's public domain.*

Participants felt that care should be taken to ensure that data was properly protected and that individuals could not be identified

4.14.2 Why share- benefits

Participants began the focus group discussion by stating that largely they were happy for their data to be shared, both because they had chosen to participate in research and did not mind, but also because sharing their data had benefits for others. These two aspects of willingness to share can be summarized by the following quotes:

> *I wouldn't be sat here if somebody hadn't helped me. (by taking part in research).*

or

> *I don't care what happens to the information, if I'm helping someone that's great.*

One participant directly dismissed altruism as the reason they took part in research, so for them, the fact that it helped others would be simply a bonus, but not the sole reason for taking part. Generally, though, participants recognized the benefits to other members of the public or to fellow patients if they took part in research and the same went for allowing their data to be shared for further research. This was especially true if the research fell into an area that was of particular interest to them:

> *…and so, the whole point of coming along is to make sure that as much is done as possible in researching the area so that we can reap the benefits.*

There was a sense that participants did not expect to benefit personally in the short term or even in their lifetime but that as long as someone was helped down the line that was acceptable, or as one participant put it, a "*no brainer*". Another stated that:

> *…I would love to think that whatever went on was passed to whoever felt that from that information they could develop things a little bit further.*

There was a high level of recognition that their data was of value to researchers. A loss of privacy was alluded to, but the compensation was that research teams were able to make their (patients') lives better.

These benefits to patients would not exist if data was not of value to the scientific community. Participants made references to *"data"* and *"researchers"* rather than to their health information, or study data specifically, pointing out that data was *"driving research"* and *"moving things forward"*. Several references were made to avoiding duplication of effort *"reinventing the wheel"* or accelerating research, saving time and building on existing research by sharing data.

4.14.3 Attitudes to sharing

Participants began the focus group largely agreeing with each other that researchers sharing their data was of no consequence to them, and that the practice was all the better if it advanced research

> *I don't care, as I said before, whatever happens isn't going to affect me personally.*

Participants were especially happy for their data to be shared if they had given permission for sharing:

> *I signed a form to say basically, that you can do what you like…*
>
> *whatever I do, I don't care what happens to the information afterwards.*
>
> *No, I haven't got a problem with that.*

However, as the discussion progressed, they were able to identify and ponder the potential consequences, intended or otherwise of sharing their data and participants began to differ

slightly in their opinions and test or question each other. Concerns about anonymity were not seen to affect them personally to any great extent, but they considered that other participants who had a disabled child, had a (rare) disease or condition, or were younger might feel differently:

> *I think if you're quite healthy your take on things is something quite different to somebody who has a disease.*

This led them to identify ways in which they might protect themselves from the consequences of data sharing and to talk about ways in which they protected their personal data in day-to-day life (although not when taking part in research), such as providing false personal details when registering for things such as email accounts and even when registering for VOICE:

> *I deliberately lie within reason. So, they can't find out about things like that (personal information).*

One participant summarized that, although data held as part of research was anonymised, someone still had to keep a record of participant's names for example, and therefore researchers still held 'personal data' for a potentially unspecified length of time. For participants personal data was name, date of birth, banking details and data about personal relationships. Disease status, however, was not considered to be 'personal' for the members of the focus group:

> *the fact that I've got an unusual disease or something... that's not privacy (personal information).*

4.14.4 The Importance of feedback

Participants were clear that after taking part in research or if their data was shared, it would be courteous if they were provided with feedback on results. Not only that but they were genuinely interested in the results. Providing participants with feedback might encourage people to take part in more research:

> *it will also make people feel quite good, well like, ok, the information that I've given is helping people in that project as well.*

They mentioned experiencing lengthy delays before they obtained feedback previously but stated that it would be still be preferable to find out how their contribution had been used

"even if it's been four years". Feedback detailing how their data had been used in research gave them *"confidence"* that researchers had listened to their views and:

> *when there is an end result…then it's quite uplifting.*

4.15 Degree of corroboration between systematic review and scoping focus group

4.15.1 Themes

The themes identified in the scoping focus group data were compared to the four identified in the systematic review prior to its update with literature published or identified after March 2018. There was a large degree of corroboration between the themes identified in the Systematic Review and the Scoping Focus group.

The benefits of sharing to patients and the participants themselves were identified by both sources, but it seemed that the systematic review participants wanted the benefits to be tangible and perhaps more apparent in the short term. This may be due to five of the papers (Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Jao *et al.*, 2015b; Merson *et al.*, 2015) coming from low and middle-income countries who may have had different expectations regarding their participation in research.

Participants in the SFG seemed to be more open to sharing, whilst at the same time having a greater understanding of the research process. Perhaps for this reason, participants in the SFG also placed slightly more emphasis upon the value of data and data sharing for researchers than those whose views were reported in the systematic review who were more focused on benefits to the community. Participants in the SFG were less concerned about security of their own data than participants in the featured papers of the systematic review, but exhibited a greater critique of security standards, and a lower degree of trust that data would stay secure.

Both groups of participants identified harms (and therefore potential barriers to them agreeing to share their data) that could occur as a result of data sharing, but the SFG participants were, at least initially, less concerned that these harms would actually occur or were less concerned about their effects. Both groups discussed different types of data and specified which types they would be more comfortable sharing. In terms of data misuse, the systematic review participants referred to accidental and intentional misuse. The participants in the scoping focus group did not refer to accidental misuse, instead focusing

upon more deliberate data manipulation, with strong emphasis upon issues of data protection and security.

Both groups discussed consent and re-consent, with agreement between groups that re-consent, although desirable in practice, was unlikely to be practicable. Participants in the studies included in the systematic review specified controls that they thought should be in place before data could be shared, whilst the SFG participants preferred to assume that controls were adequate if data was being shared between researchers or universities for example, though they became increasingly concerned about privacy and anonymisation the further away from university research the data was shared. The SFG suggested that the consent process could involve discussion of the type of organisations with whom data might be shared, and this was touched upon in the systematic review papers too, but neither group placed any time restrictions on sharing.

Both groups of participants exhibited a greater degree of trust in universities, researchers and public bodies than they did in 'commercial organisations' who were viewed by both groups as largely profit driven and more likely to use data for nefarious purposes. The systematic review participants were more certain that profit should not be made from their data. The SFG, however, identified the sometimes-necessary link between commercial research and that undertaken at universities, which had not come up in the systematic review analysis. The SFG talked little, if at all about the relationship between the researcher and participant although this was covered in more detail in several of the systematic review papers.

Overall, there was a great degree of agreement between the participants in the SFG and the systematic review findings, with many of the same issues being discussed by participants, although this could be attributed in part to the topic guide for the SFG being developed from the findings of systematic review.

It was concluded that all themes and components thereof identified by both sets of participants should be covered in the questionnaire to attempt to determine attitudes of a larger and potentially more diverse population of participants.

4.16 Question development

Following the scoping focus group, development of the questionnaire began. The intention of the questionnaire was to 'quantify' the attitudes of participants towards data sharing, for

example, to allow reporting of the percentage of respondents who had concerns about some aspect of data sharing.

As an initial source of question content, previous publications in which a questionnaire was used to collect data on this topic were identified. The relevant papers were sourced from those already known to me, including those identified during the systematic review process described in Chapter 2, regardless of whether the publication in question was ultimately included in that systematic review. Papers thus identified were only considered 'useful' if they included the questionnaire used as an appendix or supplementary information, or if they set out specific questions asked in the results section. Papers were referred to regardless of their topic; those with questionnaires about data linkage, biobank data sharing or use of GP data were considered 'in scope' as they may still have provided useful questions that could be adapted for use.

A total of eleven papers with potentially useful questions were identified and saved electronically: (Hunter *et al.*, 2009; Kaufman *et al.*, 2009; Treweek *et al.*, 2009; Willison *et al.*, 2009; Ludman *et al.*, 2010; Ahram *et al.*, 2014; Rogith *et al.*, 2014; Joly *et al.*, 2015; Patil *et al.*, 2016; Mursaleen *et al.*, 2017b; Mello *et al.*, 2018). From these papers, relevant questions were extracted and saved in a table in a Word document; where appropriate whole questionnaires were downloaded so that the questions could be reviewed. Mello *et al* (identified after publication of the initial Systematic Review) was seen as a particularly relevant example, having focussed specifically upon clinical trial participants. Questions from these identified questionnaires were not copied verbatim as often they related to sharing of biobank data rather than research data. Instead, these questionnaires were referred to, for style, wording, or ordering.

The topics covered within each theme of the scoping focus group (section 4.10 above) were also tabulated for easy reference. I noted whether these themes had been addressed as questions in existing questionnaires. If no existing relevant questions had been identified that could be drawn upon, I made a note that a new question might need to be drafted for inclusion in my questionnaire.

The themes of the scoping focus group were then incorporated into a first draft questionnaire, either in the form of new questions that I conceived myself or by adapting questions identified in the literature, for example by changing wording to place the emphasis on research data sharing or by amending or adding further response categories to

ensure that the questions reflected the scoping focus group's priorities. The first draft of the questionnaire thus developed contained 28 questions (four about taking part in research, seven about data sharing in general, four about consent, one about knowledge of sharing affecting taking part, three about storage and access, two about ownership and feedback, six on participant characteristics and one free text question). For the purposes of the questionnaire, 'storage and access' refers to placing data in a repository (or alternatively storing data with the original researchers) and how secondary researchers gain access to this data. Questions were ordered following principles for good questionnaire design of logical ordering and moving from the general to the specific. The survey began by asking whether the respondent had taken part in a trial or study (or whether their child had), how they viewed that experience, and then asked specifically about data sharing, types of consent and storage and access before concluding by gathering demographic data to allow characterisation of the sample and analysis of whether responses varied by socio-demographic status. The section on storage and access was shorter than that on attitudes to sharing more generally as the scoping focus group indicated that storage of data was the area participants were least confident about and may therefore find more difficult to answer.

Two questions about a data sharing 'register' (whereby participants who were willing to share data for all studies could be contacted) were inserted after a suggestion from an employee of Birmingham Clinical Trials Unit who contacted me by email after reading about this research in the UK Clinical Research Collaboration (UKCRC, 2021) Registered Clinical Trials Unit Network newsletter.

4.16.1 Measurements

Closed questions, with multiple-choice response formats, were chosen to minimise respondent burden. The attitudes or beliefs around data sharing needed to be measured within the questionnaire. Five-point Likert scales, with a neutral mid-point, were taken from other questionnaires such as Mello *et al* (Mello *et al.*, 2018) or devised by me. Ethnicity categories were taken from the Office for National Statistics' (ONS) suggested categories for English telephone surveys (as categories were more condensed than ONS categories proposed for other modes of response) (Office of National Statistics, 2016). Slightly narrower age categories were chosen compared to those observed in other questionnaires (using 85 and over as the highest category, rather than 65 and over) as it was felt that there could be a

difference in opinions or awareness regarding data sharing between individuals in their sixties, seventies and eighties, and that trials and health intervention studies often have participants who are aged over 65. The educational attainment category question was based upon those used by the ONS for Nomis official labour market statistics provided by ONS but adapted to fit into the limited space of a questionnaire, for example by adding 'or equivalent' after Degree (e.g.: BA, BSc) (Nomis official labour market statistics, 2014).

Survey respondents were also asked if they had any additional comments about data sharing or the survey they had just completed (a free text question). The final question asked if respondents would be interested in taking part in further research in the form of a focus group or interview about data sharing. If so, they were asked to contact me via email, with my email address placed at the end of the survey.

4.17 Questionnaire Quality

To ensure the quality of the questionnaire and reduce non-sampling errors (Biemer and Lyberg, 2003), the first draft of the intended survey was tested first through self-assessment by me, and then with cognitive interviewing (see below). Non-sampling errors refer to any error that can occur during data collection (and subsequent data processing). They include specification error, where the question as posed (and the information expected by the questionnaire designer) does not match with participants' understanding of what is being asked, and item non-response error, whereby a participant will leave a question or certain questions blank, perhaps because they are unable or unwilling to respond, or because the question is accidentally overlooked (Biemer and Lyberg, 2003). To minimise the risk of these types of error, it is important to ensure that the draft questionnaire is tested with input from the type of individuals who will be asked to complete it, to check for readability and understandability.

4.17.1 Questionnaire Self-Assessment -QAS 99

Once the first draft of the questionnaire had been reviewed by my supervisors, the Questionnaire Assessment System 99 (QAS 99) developed by Willis and Lessler (Willis and Lessler, 1999) was used to review question wording and assess the readability of the questionnaire. QAS 99 is intended to identify, and reduce as much as possible, wording or structural problems in questionnaires prior to their use *"in the field"* through a systematic appraisal process (Willis and Lessler, 1999, pp. 1-1). QAS encompasses eight dimensions of

assessment with sub questions related to: reading (of questions by interviewer), instructions, clarity, assumptions, knowledge/memory, sensitivity/bias, response categories and other, where general comments can be recorded. Each question in the questionnaire was assessed by me, in sequence, and the assessment scores (yes or no and any comments) were recorded in an Excel spreadsheet.

The first section of assessment, 'reading', was marked as not applicable for each question, as the questions would not be read out to the participant (it was designed for self-completion). For the remaining sections I erred on the side of caution as advised by the QAS guidance. Questions may not be perfect or indeed entirely fixable but having an awareness of potential misinterpretations is important.

The assessment identified several potential problem areas within the draft questionnaire (where a 'yes' answer had been given in my assessment of one or more of the eight dimensions). These were primarily at the beginning of the questionnaire; the questions intended to establish whether the participant or their child had taken part in a study, and whose data they would consider when answering the questionnaire. These problem areas identified by QAS 99 were flagged for particular attention during cognitive interviewing. No changes were made immediately, due to the subjective nature of QAS and to ensure that QAS and readability tests assessed the same questionnaire version. Instead, identified problematic sections of text were reworded as necessary post cognitive interviewing before the final version of the questionnaire was settled upon. Cognitive interviewing was conducted after QAS and readability testing and is detailed in section 4.17.3, below.

4.17.2 Readability Testing

After QAS assessment, the same draft of the questionnaire was readability tested. The sections of the questionnaire providing instructions, explanatory text or participant information were subjected to online readability tests (Scott, 2017; National Learning and Work Institute (England and Wales), 2019). Each section of text was pasted into the readability test, one at a time. The individual category responses were not readability tested as they did not meet the minimum word threshold for the readability tests. I also knew that the cognitive interviewing would help to determine the understandability of the category responses. The readability scores given to each section of text were recorded in a table. The Scott Readability Formulas use seven different readability formulas (The Flesch Reading Ease formula; The Flesch-Kincaid Grade Level; The Fog Scale; The SMOG Index; The Coleman-Liau

Index; Automated Readability Index; and Linsear Write Formula) to calculate an overall score (Scott, 2017). A description of each is given by Scott on the readability test website (Scott, 2017). The SMOG calculator (Simplified Measure of Gobbledygook) advises that a score of 14-15 is the equivalent of Adult Literacy Standard Level 2 (equivalent to GCSE grades A*-C). A score of 11-12 equates to Level 1, or GCSE Grades D-G (National Learning and Work Institute (England and Wales), 2019). The SMOG calculator is said to predict higher scores than others (Kouamé, 2010) (i.e.: worst case scenario) and so was used alongside Scott's seven formulae.

The questionnaire was then tweaked in response to the readability testing, for example by replacing words or restructuring a sentence, and then the testing was carried out again, and the table updated. Some scores indicated that readability had not improved greatly through these small tweaks and so no significant further change were made. Instead, although some of the scores were considered quite high, as care had already been taken to word the questionnaire as clearly as possible, the cognitive interviewing with VOICE (VOICE, 2017) participants (described in section 4.17.3, below) was used to further test the readability of the questionnaire. Some level of explanation (and therefore additional text) in the questionnaire was considered unavoidable, as data sharing needed to be explained to participants who might not have encountered it previously.

4.17.3 Cognitive interviewing

An application was made to VOICE (VOICE, 2017) to identify participants who would like to take part in cognitive interviews – a form of interview used to check the readability and comprehensibility of the questionnaire and to identify *"errors arising from specific stages of the response process"* (Biemer and Lyberg, 2003, p. 267). Further details on the cognitive interviewing rationale and technique are given below. An advert was placed on the VOICE website (and promoted by email to members) and 21 members indicated that they would be willing to take part in a face-to-face cognitive interview or to review the questionnaire draft by email.

I conducted face to face interviews with three members who were selected on the basis of age, gender, and experience of helping with questionnaire development. I had intended to interview four, but one of these respondents was later unavailable. I entered into email correspondence with eight additional members, who had indicated that they would like to review the questionnaire by email and received three responses which included feedback on

the questionnaire. Two additional email feedback responses were received from one of my colleagues and a member of my family. A total of eight individuals therefore gave verbal or written feedback on the draft questionnaire (see Table 6).

Participant type	Gender	Age	Experience of reviewing questionnaires	Type of contact
VOICE	Female	47	Use of questionnaires in research work. Some prior knowledge of data sharing.	Face-to-face cognitive interview
VOICE	Female	77	Looked at language in examination papers for exam boards.	Face-to-face cognitive interview
VOICE	Male	76	Assessed lay summaries (and some questionnaires) at Home Group, Newcastle University & Sunderland University.	Face-to-face cognitive interview
VOICE	Male	83	Yes	Email feedback
VOICE	Female	68	Yes	Email feedback
VOICE	Female	59	No	Email feedback
Family member	Female	60	No	Email feedback
Colleague	Female	40	No	Email feedback

Table 6: Participants who took part in questionnaire development.

Some of those expressing interest in taking part in this element of questionnaire development were not selected to take part in cognitive interviewing or email feedback as it was not considered feasible to take feedback from 21 individuals within time and budget constraints. Individuals were also rejected if their age and experience was too similar to those already selected to take part or if they wanted only to meet in person.

Cognitive Interview guides (Willis and Lessler, 1999; Willis, 2005) were referred to prior to conducting the cognitive interviews with VOICE members. The purpose of the cognitive interview is to critique or test the specific questions contained within the survey rather than the entire survey itself or the process of administering a survey (Willis, 1999, p. 7). The cognitive interviews were intended to explore how the participants understood or interpreted (comprehended) the question, whether they would decide to answer honestly and whether they thought they had the knowledge to answer and finally, whether the participant's desired response could readily be mapped to the potential given answers (for example the Likert scales used). Testing the questionnaire in this way reduces or controls the

opportunity for introduction of error or bias through participants mis-understanding questions (Biemer and Lyberg, 2003, p. 258).

Willis (Willis, 1999) describes *"verbal probing"* and *"think aloud"* interviewing techniques for use in cognitive interviewing. The pros and cons of both approaches to cognitive interviewing and types of probing are discussed in more detail by Willis in the Cognitive Interviewing Guide (Willis, 1999, pp. 4-5). Prior to undertaking the cognitive interviews, I decided to use a combination of 'verbal probing' and the 'think aloud technique' (Willis, 1999). The think aloud technique was considered useful as it requires the participant to state how they arrived at the answer to the question posed and requires very little prompting from the researcher. It was thought that this would show whether the questions were interpreted correctly and the level of difficulty that participants had in answering. Verbal probing requires more input from the researcher, but specific questions can be asked; for example, exploring how the participant has interpreted the question, by getting them to paraphrase the question back to the researcher.

Spontaneous probes (Willis, 1999, p. 9) were used as and when the need arose. However, it was judged that the most useful type of probe (to be deployed where appropriate) would be to ask the participant to paraphrase the question back to me, to ask how they arrived at an answer, to elicit whether the response options (Likert scales) provided were suitable and whether the question was easy or difficult to understand. The response categories were thought to be suitable if they provided an answer that would be relevant to the participant but also an answer that was unambiguous. It was anticipated that the conversational nature of the cognitive interviews, my relative inexperience of conducting cognitive interviews and the feedback that participants might provide un-prompted would make an entirely scripted interview inappropriate.

The cognitive interviews took place in February 2019 at Newcastle University. A copy of the questionnaire was given to the participant on the day (not in advance), and they were asked to go through the questionnaire as if they were going to fill it in 'for real', making comments where appropriate about questions that were hard to understand or could be phrased better. I held a second copy of the questionnaire which I annotated with key points during the cognitive interview. Questions that were identified as potentially problematic when assessed with QAS 99 were highlighted on my researcher copy of the questionnaire to see if interviewees also had difficulty with these questions. The scripted probes previously

identified as most useful were also printed out for researcher reference during the interview and deployed where appropriate.

The interviews were digitally recorded so that more detailed notes could be transcribed later. Recordings were deleted once transcription had taken place.

4.18 Questionnaire amendments after cognitive interviewing

Comments and responses from each participant who provided feedback (either via email or through interviewing) were recorded in a summary table, where transcribed notes were combined with those taken during the interview. Suggestions for change from participants were recorded, alongside whether or not these changes were implemented in the questionnaire, and the reason why (or why not). Changes were not necessarily made based on the views of only one participant, but change was more likely if more than one participant suggested it or if the change was obviously beneficial in terms of clarity, for example if it related to the sections identified as potential issues during QAS assessment and made the question easier to read or understand or removed superfluous words.

Comments and suggestions were received for almost every question in the questionnaire and almost every section of explanatory text. Comments ranged from those on grammar and punctuation to suggestions for improved readability and complete restructuring of the questionnaire. It was clear when multiple participants made similar comments about specific questions, which prompted change. For example, Question 9 'Below is a list of potential benefits of data sharing. Which of these make you feel more positive about data sharing?' received several comments on the structure of the statements, and ease of understanding. These statements were then amended (see Appendix E for worked examples of how comments influenced change).

Perhaps more crucially, original Questions 19-22 (questionnaire version 0.5) regarding data storage and processing were considered by two participants to be difficult to understand, and similar feedback was received from a lay member of the ACONF steering group. The questions in this section were adjusted to ease understanding and original questions 20 *'Imagine the data was stored with 'controlled access' (there is a formal request and approval process in place), where would you prefer your data to be stored prior to it being shared?'* and 21 *'Do you think there should be a limit to the number of times study data should be shared?'* were removed. Although this reduced the opportunity to learn a little about what

participants think about data storage and sharing, it was decided that it was much more important to get considered answers from fewer, better comprehended questions and to avoid participants failing to answer questions that they did not understand fully or providing responses that were invalidated by flawed understanding. Examples of changes to questionnaire text as a result of cognitive interviewing can be viewed in Appendix E.

At this stage I decided to remove Question 4, which asked whether they were thinking about their own or their child's data when completing the questionnaire and the related 'Instructions for completion' explaining this question (from the section 'Questions about taking part in research'). I judged that participants could think about whichever data they liked when completing the questionnaire and that not 'knowing' whose data they were thinking about made the data from each source more comparable at the analysis stage than if I was trying to compare participants who were thinking about their child's participation with participants who were thinking about their own participation. The questions asking whether or not participants had taken part, or had a child who had taken part, were retained for potential uses as predictor variables on strength of concern regarding data sharing. Those who had not taken part in any studies at all were still instructed to imagine that they had taken part in a research study.

4.18.1 Final readability test

After the cognitive interviewing process and removal of questions that participants found confusing, the questionnaire was left with a total of 30 questions, as detailed in Table 7.

Questionnaire Section	No. of questions
Questions about taking part in research	3
Questions about attitudes towards sharing	7
Questions about Consent	6
Questions about data sharing affecting willingness to take part	1
Questions about storage	2
Questions about ownership and feedback	2
Socio-demographics	6
Willingness to take part	1
Any comments (free text) and check box asking if they would like to take part in further research	2
Total	**30**

Table 7- Questionnaire sections and number of questions after cognitive interviewing

The readability testing was carried out again on the same sections of text and on the individual questions (but not response categories) after amendments based upon cognitive interview feedback had been made. See appendix F for examples of how the final readability

testing led to changes in questionnaire wording. The readability scores from each section of the questionnaire **after** cognitive interviewing (and therefore the final draft) are detailed in Table 8 below.

Page	Section of text in questionnaire (header title)	Scott 2017 Readability consensus readabilityformulas.com	Smog readability level (SMOG index), University of Nottingham
2	About this survey	Grade Level: 11 Reading Level: fairly difficult to read. Reader's Age: 15-17 yrs. old	17
2	What does taking part involve?	Grade Level: 10 Reading Level: fairly difficult to read. Reader's Age: 14-15 yrs. old	16.5
2	Risks and benefits	Grade Level: 11 Reading Level: difficult to read. Reader's Age: 15-17 yrs. old	19.3[7]
2	Consent	Grade Level: 9 Reading Level: fairly difficult to read. Reader's Age: 13-15 yrs. old	15.9
3	Survey background	Grade Level: 10 Reading Level: difficult to read. Reader's Age: 14-15 yrs. old	17.5
3	(Questions about taking part in research) Instructions for completion (including questions 1-3)	Grade Level: 7 Reading Level: standard / average. Reader's Age: 11-13 yrs. old	13.8
4	Questions about data sharing & Instructions for completion	Grade Level: 11 Reading Level: difficult to read. Reader's Age: 15-17 yrs. old	16.9
4-7	Questions 4- 11	Grade Level: 7 Reading Level: fairly easy to read. Reader's Age: 11-13 yrs. old	14.3
7	Questions about consent introduction (including questions 12-18)	Grade Level: 8 Reading Level: fairly easy to read. Reader's Age: 12-14 yrs. old	15.1
9	Questions about data storage introduction (including questions 19-22)	Grade Level: 7 Reading Level: standard / average. Reader's Age: 11-13 yrs. old	11
11	Thank you for taking part in this survey.	Grade Level: 9 Reading Level: fairly difficult to read. Reader's Age: 13-15 yrs. old	15.7

Table 8: Readability test results for the final draft of the questionnaire

[7] scored 17 prior to inclusion of obligatory paragraph from Newcastle University Ethics Committee.

Scores were broadly the same or had gone down slightly (indicating increased readability) after implementing amendments or suggestions for change made by cognitive interviewees. One section of text however, increased its SMOG readability score (meaning it had become less readable). This was the section on Risks and Benefits of taking part. The score increased from 17 to 19.3. This was directly due to the addition of a small paragraph of text from Newcastle University's ethics committee regarding the ethical approval given to the questionnaire. Despite a high SMOG score for this section, the readability score was in line with those for other sections of text in the questionnaire at a reader age of 15-17 years.

Although the readability testing identified areas for improvement, it was important not to rely too heavily on the readability scores to predict the understandability of the questionnaire. Readability scores are not able to measure the quality of the writing style or the overall context of the document, rather they report on aspects of the text that can be 'measured' (Redish, 2000; Kouamé, 2010), such as counting the length of sentences. For example, SMOG score calculations measure the number of words with three or more syllables (Scott, 2017), and so, although the text might be clear, if it contains many words with more than 3 syllables, for example 'participant' the associated score will still be high. Readability tests also assume that they are measuring a large body of text or traditional *"prose"* (Redish, 2000, p. 4) so their usefulness of the readability for short snippets of text (the questionnaire questions or explanatory text) can therefore be questioned, as these small sections have been taken out of the context of the document as a whole and do not form a traditional text. Finally, readability testing does not take into account things such as headings and layout which may help the reader navigate the text, or the reader's familiarity with the topic (Redish, 2000), all of which I had tried to incorporate though layout and inclusion of explanatory text.

4.19 Questionnaire Build

A licence and log in for the online survey software Qualtrics (QualtricsXM, 2021) was obtained from Newcastle University and the survey was constructed in this package. Some slight adjustments to the layout of the questionnaire were made based upon the display capabilities of the Qualtrics system. For example, where the paper questionnaire had a 'not applicable' option attached to only one question in a grid formation, the Qualtrics package had to apply a 'not applicable' option to the entire grid of questions or not at all. These

changes were then replicated in the latest draft of the paper questionnaire. A paper version was always maintained for easy reference purposes.

The survey settings were such that the survey could be distributed using a customisable anonymous link, meaning that the questionnaire did not have to be emailed to (named) respondents and I was not able to identify any individuals who had completed it. This further minimised the risk of any identification of participants but precluded the use of targeted reminders. The survey link could be customised so that, for example, the name of the group or study contacted could appear within it. The name of the study (from the link) could then be exported with the data. In this way, I could record how many responses were obtained from each group for analysis and reporting purposes.

4.20 Piloting

The electronic version of the questionnaire was then circulated to colleagues, friends, and family for piloting, attempting to further identify any functionality difficulties, spelling, grammatical errors, or difficult questions not previously identified. Eleven individuals piloted the questionnaire. Slight adjustments (for example correction of typos or spacing errors not identified previously), were made based upon the responses received during piloting. The final draft of the questionnaire as downloaded from Qualtrics can be viewed in Appendix G.

Chapter 5 Questionnaire delivery- data collection, analysis, and results

5.1 Introduction

In this chapter, I first describe the sampling strategy in terms of studies contacted and subsequently those that agreed to take part in the questionnaire survey. I then move on to describe the individual sampling frames and sample size and give a reminder of the original research question. The way in which the survey respondents were contacted is detailed, followed by the data collection, manipulation, and preparation for analysis. Any decisions made prior to analysis are outlined (for example coding of variables), with reference to the statistical analysis plan. Finally, the results of the questionnaire survey are presented.

5.2 Sampling Strategy

The target population for this survey was individuals who had taken part (or were still taking part) OR a member of the public (who could potentially take part) in public health research, clinical trials with a public health benefit or health intervention within the United Kingdom and might therefore be expected to have views on research data sharing. There is no central register or database of such individuals. Therefore, a two-stage approach to recruitment of individuals was anticipated, whereby a number of relevant studies would be identified first, followed by an approach to a sample (or indeed all) of the participants in those studies, with an invitation to complete the study questionnaire.

In the first stage, non-probability sampling was used to identify and contact appropriate trials and studies, from which study participants would then be selected. Non-probability sampling is used when particular groups appear to be representative or *"because they can be assembled conveniently"* (Fink, 2003b, p. 16). It was anticipated that participants would come from several different studies, for example a mix of trials and health interventions, whereby the participants from each might have differing characteristics and might therefore provide a diverse range of experiences and attitudes. Studies needed to have consent in place to re-contact participants about future follow-up studies. It was therefore not possible to randomly or systematically select candidate studies, e.g., from the NIHR Clinical Research Network (CRN) Portfolio database.

When checking the consent forms of trials run by Newcastle Clinical Trials Unit (my employer) I realised that very few had consent in place to re-contact participants about further studies. One exception to this was the FiCTION study, details of which are given

below in section 5.4.5. I also made enquiries with the Chief or Principal Investigators of relevant studies within the Institute of Health and Society (now Population Health Sciences Institute) at Newcastle University, with whom I was registered as a student. I made contact via email and my supervisors also sent emails or made face to face enquiries on my behalf. It was envisaged that, from each study of this type, I would be able to contact (via the original study team) only those participants who had agreed to further contact and provided contact details.

However, it became clear from investigator responses that many studies lacked the explicit consent or ethical approval required to re-contact participants regarding either follow-ups to the original study or participation in similar future studies. In addition, and to a lesser extent, some investigators did not reply in a timely manner or had plans themselves to re-contact participants with invitations for follow-up studies related to the original, and understandably did not wish to over-burden participants with invitations to take part in other work.

The search for a sample of studies from which I could draw participants was then extended to outside of my home institute and workplace, for example to organisations with links to Newcastle University such as Fuse (Fuse, 2022). I utilised as many contacts within and outside of the university as possible and approached Chief/Principal Investigators of clinical trials and longitudinal studies, as well as public health research, by email. Investigators of relevant NIHR public health studies were contacted by email on my behalf by a former colleague. Lists of NIHR funded Programme Grants for Applied Research and Research for Patient Benefit studies were searched online, with the corresponding chief or co-investigators emailed by me. CHAIN (CHAIN, 2022) members and UKCRC registered trials unit members were contacted on my behalf after I submitted written requests to the respective networks. by institutions and organisations contacted as detailed in Table 9 below.

The sampling strategy had therefore moved more towards a sample of convenience; or *"a group of individuals that is ready and available"* (Fink, 2003a, p. 18), or participants who had taken part in a study which had ethical approval to re-contact them and where there were no practical or financial implications that could not be overcome. Oppenheim describes a *"judgement sample"*, a sample where *"accurate parameters for the population are lacking"* but researchers have done their best to ensure that the sample contains as diverse a sample of individuals as possible, which is what I tried to do (Oppenheim, 1992, p. 43). This type of sampling may also be referred to as *"purposive"* (Biemer and Lyberg, 2003). A balance had to

be sought between participants who I would be able to contact, and the practicalities of doing so, or as Oppenheim stated: *"compromises between theoretical sampling requirements and practical limitations such as time and costs"* are often required (Oppenheim, 1992, p. 43).

The advantage of using a convenience or judgement sample in securing participation only from studies where investigators allowed contact of participants is the low or reduced cost. The broad target population for my study meant that there were no stringent inclusion criteria that could preclude certain studies being included in this research, meaning many types of study (e.g., longitudinal cohort studies) could be included, as long as they had a public health benefit or health intervention angle. Studies that were not health related at all were still not considered.

Obtaining responses from as many different types of study as possible was the aim, so that if possible, the sample should be sufficiently large and diverse that generalisability could be argued, albeit with caution. It is not possible to state that the attainable sample of participants would be representative of the general population of the United Kingdom, but that was not the intent in any case. It was plausible, however, that the achieved sample would be broadly representative of participants in each of their respective studies, and perhaps of research participants in public health (intervention) studies and clinical trials in general. Obtaining responses from participants from a variety of studies run by different institutions, and therefore based in different geographical locations, increases this sense of cautious generalisability.

5.3 Inclusion criteria for participants:

The specific inclusion criteria for the participants selected to take part in the questionnaire survey were as follows:

- Aged over 18
- Capacity to give informed consent to take part;
- Resident in the UK (or taking part in a study that originated in the UK, if current place of residence unknown);
- Currently taking part or had taken part (or their child had taken part) in a health research study, longitudinal study or clinical trial OR a member of the public (as a potential participant) with particular interest in research studies.

Where the source studies had very large numbers of participants, a focus on particular sub-samples of study participants was discussed on an individual basis with the investigators or study teams concerned.

5.4 Studies contacted

Contact was made with the following types of study/organisation listed in Table 9 below. Those which were successful are presented separately and highlighted in blue.

Contacted	**Outcome**
Fuse (The Centre for Translational Research in Public Health)	No studies able to help due to lack of consent or ethical approval for re-contact.
VOICE (PPI allowing participants to take part in and contribute to research)	Application made, and survey link sent to ~3000 members. Interested members completed survey.
Investigators within the Institute of Health and Society, Newcastle University, plus Investigators contacted by supervisors (n=11)	One response but no permission to contact participants in place.
NIHR Public Health Studies (n= 7)	1= replied but not able to help 6= no response
UKCRC Registered Trials Units newsletter recipients (Clinical Trials Units)	1 interested response but making survey content suggestions only.
Longitudinal Studies contacted individually e.g., Gateshead Millennium Study, Born in Bradford Study, Thousand Families Study, Southampton Women's Study (n= 24)	4= sent invitations for applications or entered further discussion (all unsuccessful) 10= no response 10= replied but not able to help- no permissions or participant burden considerations.
Centre for Longitudinal Studies (requesting access to 1970 British Cohort Study, The Millennium Cohort & The 1958 National Child Development Study).	Positive response to formal application but application declined due to anticipated patient Burden. Future collaboration encouraged.
Aberdeen Children of the 1950s study	Application successful and questionnaire distributed to 1400 participants.
ALSPAC	Application successful and questionnaire distributed to 5858 participants.

Contacted	Outcome
NIHR Research for Patient Benefit Studies (RfPB) (n=5)	1= replied but not able to help 4= no response
NIHR Programme Grants for Applied Research (PGfAR) (n=8)	1 = response encouraging of research but not able to help due to resource issues 2 = replied but not able to help 5 = no response
SAIL Consumer Panel and SUPER Group (PPI group allowing participants to take part in and contribute to research) (combined members n=30)	12 responses to questionnaire.
FiCTION Trial, Newcastle Clinical Trials Unit (NCTU) (FiCTION futures)	Application to Chief Investigators successful, but NCTU required original REC approval. This was also successful, but COVID-19 pandemic prevented physical access to participants' contact details.
CHAIN (Contact Help Advice and Information Network for people working in health and Social Care). Email sent to members.	1 response giving advice only.
TOTAL contacts made	64

Table 9- Studies and Organisations contacted to identify participants for the questionnaire survey.

Chief/principal investigators of suitable studies were contacted to determine; firstly, if they had permission to re-contact their study participants for further research; and secondly, if they would be willing to do so, providing an invitation to complete my questionnaire. Some investigators, such as those within Newcastle University, were initially contacted directly via email by myself or my supervisors and followed up (by email, phone, or in-person) if they were not initially responsive. Responses were saved and categorised into yes and no responses. Any positive responses were followed up with informal email discussion and if necessary, by formal applications for participant contact as required by the study in question.

Of the investigators who responded (n=29), many did not have ethical approval to re-contact participants, had no means of re-contacting participants or did not wish to over-burden them with further research topics (n=16) (in some instances, the sentiment of 'saving' the pool of participants for future follow-up by the original researchers was expressed). Some

responded with encouragement and a favourable opinion of this research study but were not able to help (n=3) and some offered future collaboration (n=1). Five responses got as far as formal applications that were then rejected due to avoidance of participant burden, or email discussions which did not result in collaboration. Some (n=35) did not respond at all, even if followed up.

Despite the initial difficulty in obtaining permission to contact study participants, three investigators agreed that their participants would be suitable for contact. These studies were: The Aberdeen Children of the 1950s longitudinal study, the FiCTION Futures study and the Avon Longitudinal Study of Parents and Children (ALSPAC); the latter agreed to allow a sub-group of their Children of the 90s participants to be contacted with an invitation to take part in my survey. Three appeals to PPI groups (VOICE, SAIL and SUPER) were also successful. Further details on the included studies follows below.

5.4.1 Aberdeen Children of the 1950s

This cohort is made up of 12,150 participants (6276 males, 5874 females) born in Aberdeen between 1950 and 1956, who took part in the Aberdeen Child Development Survey, completed in local primary schools in 1962. Data collected include information on *"birth weight, childhood height and weight, tests of cognition and behavioural disorder, and a range of multi-level socio-economic indicators*" (Batty *et al.*, 2004). There have been various follow-up and sub- studies since 1962, but in 1998 the study was 'revitalized' and is now referred to as the Aberdeen Children of the 1950s Study (ACONF). As of 2004, the location of 98.5% of the original participants was known with 81% still resident in Scotland, 73% still resident in the Grampian region and 500 known to have died (Batty *et al.*, 2004). Linkages to hospital admissions and other health outcomes available through the routine data have been made, and a 1998 postal questionnaire to all surviving cohort members obtained a response rate of 64% (Batty *et al.*, 2004).

After an informal email enquiry, requesting permission to contact the participants of the study, I was invited to apply to the study steering group, providing them with a short study protocol detailing my participant requirements, evidence of ethical approval and study rationale. The steering group reported back that, prior to approval, they wished to view the list of questions that I intended to use in the survey, to ensure that participants were not answering questions that had already been explored with them at workshops or public engagement events. There was also a requirement that cohort-specific results compared to

other groups and general results be shared with the Aberdeen Children of the 1950s study team on conclusion of my study. One member of the steering group commented upon the questionnaire stating that it was *"awfully long and complicated… particularly towards the end – when it presumes an intuitive understanding of issues to do with storage etc (e.g., Q19-Q22)"*. This section of the survey regarding storage was amended and made more concise in light of this comment and similar feedback received during cognitive interviewing (see Chapter 4).

It was initially agreed that I would contact approximately 1000 participants, (equal gender split), and administrative costs were set out. Later, a brief invitation letter was drafted for the study participants and reviewed by the Study Manager. The Study Manager then contacted me to provide the survey link and it was sent to 1400 participants who were registered to receive a mass email on 4th October 2019. The study did not invoice for the administrative work involved in contacting these participants.

The survey was available for participants from 4th October 2019 until 7th January 2020.

5.4.2 Avon Longitudinal Study of Parents and Children (ALSPAC)

Avon Longitudinal Study of Parents and Children (ALSPAC) or Children of the 90s (Boyd *et al.*, 2013) recruited pregnant women in the Bristol area between 1990 and 1992, resulting in recruitment of 14,541 pregnancies (G1). ALSPAC also collects data on the parents of the pregnancies (G0) and resulting offspring of the original 1990s cohort (G2). Data collected includes phenotypic and environmental measures, biological and genetic/epigenetic samples, linkage to health and administrative records and questionnaire data.

ALSPAC welcome requests from researchers for data, samples or the opportunity to collect new data. After informal discussions with the ALSPAC team, an application form was submitted to ALSPAC in June 2019. This application was subsequently given approval by the ALSPAC executive two weeks later. However, it was not until November 2019 that costs and the method of contacting ALSPAC participants was confirmed, as the study team were busy sending an annual questionnaire to participants.

ALSPAC also have their own participant panel, termed the original cohort advisory panel (OCAP), who needed to review the documentation associated with the survey (survey questions, PIS and Invitation letter) and an ethics committee (ALEC), who required a separate application form to be completed, despite approval from Newcastle University

Faculty of Medical Sciences (FMS) Research Ethics Committee (REC) already being in place. The OCAP made a few minor suggestions for changes to the questionnaire, including removal of the question asking for postcode (included to allow for calculation of index of multiple deprivation). To facilitate a smooth subsequent ethical review by the ALEC, it was decided that postcode would not be collected for ALSPAC participants. The ethics committee (ALEC) made a more comprehensive list of suggestions for amendments to the questionnaire, and these were responded to formally with a new draft of the questionnaire sent back to ALSPAC. The amendments, although multiple, were minor in nature and concerned wording, for example making it clear that the questions regarding data sharing were hypothetical and would not change the processes by which ALSPAC handle participant data.

Once this process had been completed, the survey was available for members to complete in May 2020.

I had requested that the original cohort (G1) were contacted, since, with their years of birth being in 1990-92 (Children of the 90s), they were a younger cohort than can often be found in clinical trials or health intervention studies, and of a different generation to the ACONF cohort. There were approximately 6000 G1 participants eligible to be contacted by email, but those who are flagged as deceased, withdrawn, or who said no to questionnaires and no to contact were excluded by ALSPAC staff. Further to this, participants who require additional management to complete questionnaires (termed 'safeguarding' by ALSPAC) were reviewed on a case-by-case basis by the participation team and a decision was taken on whether it is appropriate to include them. The final number contacted was 5,858. The survey was distributed in batches of 200 over a period of 2 or 3 days.

There was a cost involved in contacting the ALSPAC participants which was negotiated with the ALSPAC survey team and kindly paid by my main supervisor from her research account. To keep costs down it was agreed that only those participants who had opted to receive surveys via email would be contacted, and there would be no reward/incentive for taking part or no reminders sent if the survey was not completed following initial contact (both of which are usual for ALSPAC surveys). The survey was however publicised using social media run by the ALSPAC team. It was recognised that the lack of incentives and reminders would be likely to have a detrimental effect on response rates (Edwards *et al.*, 2009), but resource constraints required balancing costs against quantity of response.

The ALSPAC team preferred that the survey was set up in REDCap (their standard collection tool) in the same format as their usual study surveys, rather than in QUALTRICS as for the other participants. To access the questionnaire each participant is given a unique security token in addition to a username/password. The link emailed to the participant contains their specific security token. When the participant clicks on the link, it maps the security token to the participant and logs them in. This allows access to the questionnaire and for data to be submitted.

This meant that the survey data for ALSPAC participants was collected separately to that of other participants, and then sent to me to check and merge with the data of the other studies. ALSPAC staff also reviewed the responses to the free text question number 29 and redacted as necessary to ensure that the contents did not allow identification of the participant in any way, before the data, including the free text response was sent to me.

As part of the ALSPAC Access Policy researchers must agree not to share the data with anyone not named on the application (including data sharing of anonymised data with other researchers), to destroy the data at an agreed time point and to return generated variables to ALSPAC. It was agreed that the data would only be held until the end of the study (the last possible date at which I would be likely to receive questions about any publications or be making corrections to the thesis), meaning that all analysis and publications had to be complete before the data was destroyed.

The survey was made available to the first 200 participants on 14th May 2020, and responses were monitored by the ALSPAC team until they began to tail off. The survey was closed on 7th July 2020.

5.4.3 VOICE

To gain further insights into different groups of participants, the survey link was also sent to members of VOICE (formerly VOICE North) (VOICE, 2017). VOICE is a patient and public involvement (PPI) group based in the UK National Innovation Centre for Ageing (NICA), and was founded in 2007. Although VOICE is based in the North East of England, its members are spread geographically throughout the UK. VOICE members provide patient and public involvement, in the form of input, ideas and feedback into research activity, but also to businesses, charities and community members, helping to shape products and services (VOICE, 2017). VOICE members are primarily, but not exclusively older adults, with an

interest in, but not necessarily direct experience of, taking part in research. The VOICE website has a formal application process for researchers, and once projects are approved, an advert is sent to members by email as part of a weekly newsletter. Links to research projects also stay active on the VOICE website until the research end date is reached. Researchers who use VOICE participants are required to provide timely feedback to members on the ways in which their participation impacted upon the research study. This feedback is also published on the VOICE website and distributed to members via email.

I had previously contacted VOICE to determine how many members they had on their register, to inform estimates of response rate. In January 2020 VOICE announced that they had joined with Imperial College London, meaning that a wider pool of participants could be reached with one survey invitation. VOICE members include researchers, members of the public and representatives of other interest or research groups. Excluding members who choose not to receive 'invitation to take part' mailers or the newsletter, there are approximately 3000 individuals on the VOICE mailing list.

An advert was placed on the VOICE website and distributed to members via the weekly newsletters. The same 'invitation to take part' text as used for the invitations to ACONF and ALSPAC participants was used and accompanied the survey link in the VOICE newsletter (see Appendix H). Members were able to click on the survey link and be taken straight to the survey introductory page in a new tab.

The questionnaire was available for completion for the months of December 2019 to the end of June 2020.

Feedback in the form of a summary of PPI group members' results was provided to VOICE in January 2021 which was then distributed to members who took part. The same summary will also be published on the Voice website as a blog post but at the time of writing (December 2021) it had not yet been posted.

5.4.4 SAIL Consumer Panel and SUPER Group

Through my work in the UK Registered Clinical Trial Units (UKCRC) data sharing group, I was able to contact the patient and public involvement (PPI) representative of this group and ask if they would be willing to complete my survey and also if they knew of any other PPI groups who might also be willing. The group's PPI representative forwarded my survey to the SAIL

(SAIL Databank, 2020) and SUPER (PRIME Centre Wales, 2018) group members, of which there were 30 in total.

SUPER (Service Users for Primary and Emergency care Research) group members are from diverse backgrounds within Wales, recruited by PRIME Centre Wales (http://www.primecentre.wales/) to support and give patient and public perspectives on research activity, in particular research development and dissemination. The research activity is, as the name suggests, focussed upon primary and emergency care.

The SAIL Databank PPI group (SAIL consumer panel) was established in 2011 to provide the public's perspective on research into data linkage in areas such as safeguarding and ethical approval, and to provide input on projects from bid to approval and dissemination stage. Members are involved in *"all levels of the management of SAIL databank"* (SAIL Databank, 2020). There are also opportunities for members to join the teams of individual studies.

The survey was sent to users of both groups by anonymous link in an email from the UKCRC data sharing PPI member.

The survey was available for completion by SAIL and SUPER group from 7th October 2019 until June 2020.

5.4.5 FiCTION

The FiCTION (Fillings in Children's Teeth, Indicated or Not) dental trial (Innes *et al.*, 2013) was a 3 arm multi-centre randomised trial comparing three treatment strategies (conventional, biological, prevention) applied to children with caries aged 3-7 years old, over a period of 3 years.

The FiCTION trial was managed by Newcastle Clinical Trials Unit (NCTU) and sponsored by the University of Dundee. The Chief Investigators were contacted directly to see if they would agree to allow parents of FiCTION child participants who had given prior permission to be approached about follow up-studies to be contacted, inviting them to take part in this questionnaire. This group of parents was termed the FiCTION Futures group. The Chief investigators were encouraging and granted permission for the relevant parents to be contacted. Parents who had consented to be approached about further studies had provided contact details (including email address) in a contact form which was stored securely at NCTU.

NCTU required permission from the original ethics committee prior to allowing access to the participant contact details. This permission was granted by email after some liaison. The original ethics committee also suggested that the research manager of Newcastle University's Faculty of Medical Sciences was also contacted for permission to contact the participants. This permission was subsequently granted, but this process took time. All permissions were in place shortly prior to declaration by WHO in March 2020 of a global Coronavirus pandemic and move to home working by all University staff. This led to a delay in access to the NCTU building, and ultimately the FiCTION participants were not able to be contacted in time to collect data and combine the results with the other datasets. FiCTION participants were therefore not included in this study. This study is mentioned here due to the successful outcome of contact with the investigators and the time spent liaising with the investigators and NCTU.

5.5 Sample Size

Once the sampling strategy had been identified, the size of the achievable sample had to be calculated. Sample size was largely determined by the number of studies that agreed to take part, the number of participants that could be contacted for each study and the survey response rate.

As indicated above, a total of 10,288 individuals were sent questionnaires. A range of factors, both modifiable by the researcher and outside of researcher control, have been identified as affecting overall response rate to surveys (Edwards *et al.*, 2009). These include the topic of the survey, whether the survey is unsolicited and/or from an individual or organisation already known to the respondent, the length and complexity of the questionnaire, whether there is prenotification of the arrival of the questionnaire, the number and nature of reminders, and the provision of incentives for response (McColl *et al.*, 2001; Edwards *et al.*, 2009). The stipulations of the investigators of the source studies, the mode(s) of contacting potential respondents and the limited resources available for this study precluded the use of pre-notification, reminders and incentives. As already recognised, it was accepted that this would be likely to lead to relatively low response rates.

Fink (Fink, 2003b) suggests that *"unsolicited"* surveys receive the lowest response rate with around 20% being common (Fink, 2003b, p. 56). A typical response rate of *"up to 25 percent"* for web surveys with prior mail invitations in Slovenia was reported by Vehovar and Bullens,

but so was a response rate as low as 10% for a survey with just one *"initial contact"* with rates increasing up to 40% with 3 reminders sent (Vehovar and Beullens, 2018, pp. 34-37). A systematic review of web-based surveys distributed by email in the 1990s reported response rates of between 6% and 63% (Schonlau *et al.*, 2002). A previous study of ALSPAC cohort participants (Bray *et al.*, 2017) found that participants were 10% less likely to respond if sent an invitation to a web survey alone than they were if they were offered a choice of responding online or by post. Other recent surveys completed as part of trials with text messaging reminders found high completion rates of up to 97% for postal questionnaires (Keding *et al.*, 2016; Cochrane *et al.*, 2020). Conversely, de Vaus suggests a larger non-response rate of 30% (and resultant response rate of 70%) if the best follow up techniques are used. Given that reminders and other optimal follow-up techniques could not be followed due to budgetary constraints, I considered the de Vaus estimate overly optimistic.

Based on the response rates reported in the literature above, previous (unpublished) undergraduate survey work I had carried out, and discussions with the individual study teams prior to questionnaire distribution, I therefore cautiously assumed an overall survey response rate of 20%. Given that the number of participants contacted and invited to take part totalled 10,288, I therefore expected an achieved sample of approximately 2,058 (1,172 from ALSPAC, 280 for ACONF and 606 from VOICE, SAIL and SUPER combined).

The nature of the data collected in this survey was such that an appropriate summary measure was the percentage or proportion of respondents holding a particular view regarding data sharing, e.g., being 'concerned' if they knew that data from the study in which they were involved was being shared. The formula for calculating the target achieved sample size to estimating a proportion (p) to within a given margin of error (d), with 95% or 99% confidence, is given by: -

$N = p \times (1 - p) \times z^2 / d^2$

Where:

p = the estimated proportion in the underlying population

z = 1.96 for 95% confidence; 2.58 for 99% confidence

and d = the acceptable margin of error (also expressed as a proportion).

The confidence interval represents the score or figure that would be obtained if the survey were to be repeated many times over using different participants. The narrower the confidence interval, the more accurate the survey (Meterko *et al.*, 2015). The larger the sample, the narrower the confidence interval around parameter estimates or *"sampling variation"* (Fink, 2003b, p. 29).

Setting p at 0.5 provides the most conservative estimate. For 95% confidence, and a margin of error of ±5%, the required achieved sample size is 385; for 99% confidence and the same margin of error, it increases to 664 (Dhand, 2014). If higher precision is required, i.e., if the margin of error is reduced to ±3%, sample size increases to 1068 or 1,849 for 95% or 99% confidence respectively. It was therefore judged that the anticipated sample size was likely to yield adequate precision for estimates based on the combined responses from all study sources, even if the overall survey response rate fell somewhat short of the expected 20% and/or in the face of item non-response on key variables.

As indicated below, cross-tabulations and the chi-squared statistic were used to identify any associations between the dependent variables of attitudes towards and preferences for data sharing, and nine independent variables. Bujang *et al* have suggested that a minimum sample size of 500 for observational studies of large populations (Bujang *et al.*, 2018). They also cite a 'rule of thumb' for determining sample size for this type of analysis of 100 + 50 x i, where i is the number of independent variables; for this study, this would suggest a minimum sample size of 550 (Bujang *et al.*, 2018, p. 126). The anticipated overall achieved sample size of 2,058 comfortably exceeds these thresholds, suggesting adequate power and precision should be achievable.

5.6 Questionnaire distribution

With the exception of the ALSPAC study (detailed above), the questionnaire was distributed through a Qualtrics 'anonymous link' provided to participants by their original study team or contact. No personally identifiable data was collected from participants in the questionnaire, and I did not need to know contact details (e.g., email address) to be able to distribute the survey. I was also able to send a separate link to each study involved incorporating a word in the link as a marker which could be downloaded telling me which study the participant came from, e.g., 'source'=ACONFAC.

When first embarking upon this PhD study I had anticipated that I would also send paper copies of the questionnaire to respondents who would prefer to complete them this way, but by the distribution stage it became clear that this would not be feasible. I did not have access to the contact details of the participants who took part from ACONF or ALSPAC. I chose not to ask the study administrators from ACONF or ALSPAC to send paper questionnaires to participants on my behalf as this would have attracted additional administrative costs. I then chose not to give VOICE, SAIL or SUPER group participants the option to complete the survey on paper to prevent me seeing participant contact details, to avoid the administrative costs to myself, to ensure that all respondents had equal opportunity to complete the questionnaire and to remove the risk of mode of administration effects.

5.7 Research question

Broadly the analysis attempted to determine participants' attitudes towards data sharing.

The specific research questions (RQ) of the PhD study were:

1. What are participants' attitudes towards data sharing (and how may these differ according to socio-demographic characteristics and prior research experience)?
2. Does knowing about it affect their likelihood to participate in research?
3. What are their preferences regarding data sharing?
4. To what extent does current guidance reflect research participants' views and priorities?

There were no pre-determined hypotheses that needed to be tested, instead the analyses were largely exploratory, and descriptive within the scope of the specific PhD study questions. The questionnaire analyses attempted to answer (and expand upon where possible) research questions 1-3 above. Table 10 below details the questionnaire items used to address each of the three research questions above.

	Research question	Sub questions	Questions used to answer this	Sub-items by question
RQ1.	Attitudes towards sharing	How concerned	Q5	-
			Q6	Q6_1 - Q6_9
			Q7	Q7_1 – Q7_8
			Q7a	-
		How likely to give permission	Q8	Q8_1 – Q8_6
		Things that encourage sharing	Q9	Q9_1 – Q9_5
			Q10	Q10_1- Q10_5
		Which data to share	Q11	Q11_1 – Q11_15
RQ2.	Does knowing about data sharing affect the likelihood of respondents taking part in research?		Q15	-
RQ3.	What are respondent's preferences for sharing?	Consent	Q12	Q12_1 – Q12_5
			Q13	Q13_1 – Q13_6
			Q14	-
			Q16	-
		Storage	Q19	-
			Q20	Q20_1 – Q20_6
		Register	Q17	-
			Q18	-
		Ownership	Q21	Q21_1 – Q21_6
		Feedback	Q22	-

Table 10- Research Questions and how they are answered by the questionnaire

Independent variables characterising the respondents were analysed alongside the dependent variables above to see if responses were answered differently depending on respondent type – for example, whether or not respondents had taken part in research previously, the type of study they took part in (data source), or demographic characteristics.

The independent or descriptive variables were as follows:

- Age
- Gender
- Education
- Ethnicity
- Overall health at the time of response
- Deprivation (using Townsend/Carstairs score determined by postcode)
- Source by which the respondent was identified
- Whether the respondent had personally taken part in a research study
- Whether the respondent's child had taken part in a research study
- The respondent's personal experience of taking part in a study (positive or negative)
- The respondent's child's experience of taking part in a study (positive or negative)

5.8 DATA CLEANING AND MANIPULATION

An analysis plan was developed to describe how the data would be used to answer the research questions outlined in the introductory chapter (Chapter 1), how the data would be prepared for analysis and which analysis techniques would be used. This information is summarised below.

For data collected through Qualtrics, survey responses were downloaded as .csv files using the inbuilt 'Source' marker as a variable to differentiate one study's respondents from another. The Qualtrics system does collect IP addresses of survey respondents, but these were immediately deleted from the output data to preserve confidentiality. The data from ALSPAC respondents was sent by ALSPAC themselves, as a fully labelled Stata file. Both data sets were imported into and merged in the statistical software package StataIC (version 15). Questions from ALSPAC needed to be re-numbered to match the remainder of the data.

5.8.1 Cleaning

Very little cleaning was required for this data set. Due to the way in which data were collected, missing answers or answers that seem contradictory could not be checked and corrected. The questionnaires in both Qualtrics and REDCap (for the ALSPAC collection) were constructed to minimise both errors in data collection and missing item responses. This was achieved by ensuring that the questionnaire layout was as clear as possible, brief instructions were included, the questions themselves were clear and made sense to respondents (see Chapter 4), and that questions had adequate and meaningful response categories including 'not sure' where appropriate (de Leeuw, 2001).

There were no mandatory questions in the Qualtrics or REDCap questionnaire builds, so it was possible for respondents to skip questions. It was noted which specific questions had a large amount of missing data. No imputation of missing values took place. Given the number of responses received, it was thought that it would be unlikely that missing data would have a significant impact upon the analysis or that patterns of missingness would be informative.

Data was 'sense checked' for example, if respondents answered for question 13a that nothing would convince them to share their data, a check was made to make sure they did not also provide responses to other sub-parts of question 13 regarding factors that would encourage them to share their data. For those that had (n=9), a sensible correction was made by removing the answer to Q13a. Any corrections of this nature were documented.

Adjustments to the data after download from Qualtrics or received from ALSPAC (e.g., removal of additional headers) before data was imported into Stata were documented so as to be replicable. The analysis plan detailed any anticipated cleaning that needed to be performed prior to analysis.

5.8.2 Free text responses

The responses to question 29- 'Do you have any further comments about data sharing or about this survey?' were summarised separately to the quantitative data and responses such as 'no' or 'none' were removed. ALSPAC had removed 3 potentially identifying words or phrases and replaced them with the word 'REDACTED'. Remaining responses were read and categorised (or coded) based on the primary emphasis or *"polarity"* (Richards *et al.*, 2009) of the comment. Comments were then re-read, and categorisations were adjusted as necessary. During this process categories were amalgamated, or new categories were identified as described by Cunningham and Wells (Cunningham and Wells, 2017). Finally, categories were summarised quantitatively.

5.8.3 Checking for missing data

At ALSPAC's request, ALSPAC respondents did not have an explicit 'prefer not to say' option on the demographic questions; instead, if ALSPAC respondents preferred not to give this information they had to leave these questions blank. ALSPAC respondents were also not asked whether they took part in a study, or whether their child took part in a study, as the answers to these questions are already known in principle (though we cannot know if the respondent had a child who has taken part in a study outside of ALSPAC). The answer to Question 1, 'Have any of the following ever taken part in a health research study?' was therefore be presumed to be 'you', although no answers were input on behalf of ALSPAC participants where they did not exist.

5.8.4 Deprivation score calculation

Postcode was requested so that a deprivation score could be calculated for respondents. At the request of the ALSPAC study group, no postcode data was requested from Children of the 90s respondents so a deprivation quintile could not be calculated for this group. In total the postcode question was only answered by three hundred individuals or about 65% of respondents from ACONF and the PPI groups. These respondents provided full or partial postcode data, and only 4 postcodes were unable to be determined from partial postcodes

given, meaning a deprivation quintile was calculated for 296 respondents or 17.6% of total respondents.

Where postcode was available, the respondents' deprivation score was calculated using either Townsend (UK Data Service, 2020) (for participants in England) or Carstairs score (Brown *et al.*, 2014) (for participants in Scotland). Deprivation scores were obtained using various look up files freely available on the internet and the V LOOK UP function in Microsoft Excel. To obtain the Townsend scores, postcodes had to be converted to ward name. For partial postcodes, the most likely ward was assigned based on available digits of postcode.

Townsend scores indicate quintile 1 as most affluent and quintile 5 as most deprived. From 2011 Carstairs scores were changed, where quintile 1 indicates most deprived and quintile 5 indicates least deprived. It would have been too confusing to compare scores which use the same scale but in the opposite direction, and so the Townsend scores (as there were fewer (n=69) respondents from England who gave postcode) were transformed as detailed in Table 11 so that the Carstairs and Townsend scores matched.

Original Townsend Score	New Score given	Carstairs Score
1 most affluent	5	1 most deprived
2	4	2
3	3	3
4	2	4
5 most deprived	1	5 most affluent

Table 11- Transformation of Townsend score for analysis

Deprivation is therefore reported as a (transformed Townsend or Carstairs) quintile (1-5), where 1= most deprived, 5= most affluent.

5.8.5 Respondent groups

Prior to summarising data and performing analysis, the dataset was manipulated in Stata so that there were three respondent groups:

- PPI groups (comprised of SAIL, SUPER and VOICE data combined),
- Aberdeen children of the 50s (ACONF); and
- ALSPAC.

For the purposes of analysis only I had intended to create a new variable further dichotomising the respondent groups by combining those currently taking part in a research cohort study (ALSPAC and Aberdeen children of the 50s) and those who are members of PPI groups (VOICE, SAIL, SUPER group). However, the number of participants who were taking part in a PPI group was too small (n=63) to meaningfully compare to those in ALSPAC and ACONF (n=1,621) in analysis.

5.8.6 Recoding

All respondents (PPI groups and ACONF) other than ALSPAC were given the option to answer 'prefer not to say' to questions 23 to 27, the demographic questions that made up the bulk of the independent variables. The number of respondents who selected 'prefer not to say' was minimal, compared to the number of respondents who chose to leave these questions blank if they did not want to provide their demographic details. Table 12, below, details the number of respondents who chose to answer 'prefer not to say' as opposed to leaving the question blank.

	Source study		
Question	Aberdeen	PPI groups	Total
Q23 gender	0	1	1
Q24 Age	1	3	4
Q25 Ethnicity	2	3	5
Q26 Education	8	2	10
Q27 Overall health	0	1	1

Table 12- Respondents selecting 'prefer not to say'.

As so few respondents had answered in this way, and because ALSPAC participants did not have this response option I decided, therefore, to treat 'prefer not to say' as missing and recoded these responses as such.

Data summaries also revealed that there was just one respondent each in the age groups 18-24 and 85 and over. For the purposes of secondary analyses only, these answers were re-coded so that the respondent aged 18-24 moved to the 25-44 category and the 85+ respondent moved to the 75-84 category.

5.8.7 Variable dichotomisation

Although Likert scales are useful for measurement and for descriptive analyses, they do tend to assume that attitudes are scalable and linear (Oppenheim, 1992, p. 200). This assumption is not always warranted. We cannot be sure that respondents see the difference between 'very concerned' and 'somewhat concerned' as equal to that between 'somewhat concerned' and 'not at all concerned'. It was therefore decided not to treat the responses to the dependent variable items as continuous data (interval scale) in the secondary analyses. Treating each question as having binary responses allows for uniformity in the analysis.

To perform secondary analyses, any categorical (Likert scale) response variables therefore needed to be collapsed into dichotomous variables, for example a positive view versus a negative view of data sharing. A new copy of the data was saved, and this data was dichotomised as necessary in Stata. The 'not sure' options in the Likert scale also had to be incorporated into the dichotomous categories for the purpose of analysis. Each question with 'not sure' as a response option was considered individually, depending on where the question's emphasis lay. It was decided to err on the side of caution by assuming, for example, that respondents answering 'not sure' were doing so because they were more cautious about sharing rather than unequivocally unconcerned or wholly ambivalent. Once survey responses were received it was observed that relatively few respondents used the 'not sure' options, so re-categorizing such responses as either positive or negative response was not thought likely to have a significant biasing effect on the analysis. Table 13 below gives an example of how Question 5 was split so that it had a binary outcome. Full details of the dichotomisation of the variables are given in Appendix I.

Dependent variable	Questionnaire answers	Dichotomous answers	
Q5 How concerned would you be if you knew data from the study that you are involved in was being shared?	• very concerned • somewhat concerned • not very concerned • not at all concerned • not sure • depends who it is shared with	Concerned	• very concerned • somewhat concerned • depends who it is shared with • Not sure
		Not concerned	• not very concerned • not at all concerned

Table 13- Example of a dependent variable split into binary responses.

5.8.8 Other checks prior to analysis

Questionnaire data, particularly those using Likert type scales, are not continuous and are not necessarily going to be evenly distributed across all response categories; they are ordinal, and the difference between each interval (or point on the scale) is not necessarily equal in measure or indeed measurable (Sullivan and Artino, 2013; Cooper and Johnson, 2016, p. 175). It is possible that most respondents will choose the same response e.g., 'very concerned' which would skew the data to one end of the Likert scale, resulting in a non-normal distribution. Production of histograms in Stata identified that for most variables, the answers were skewed towards the positive end of the scale; where respondents are either positive about sharing data or have areas of concern.

5.9 STATISTICAL METHODS

This section presents the statistical methods used in summaries and analyses of the survey data and the justification thereof, including variables that were excluded, and significance levels. Data summaries were produced, followed by cross-tabulations of independent variables and production of cross-tabulations or contingency tables including the results of a Chi square test for all dependent variables with each independent variable in turn.

5.9.1 Missing data

Incomplete questionnaires were included in all summaries and analyses. Respondents with missing answers were excluded from analysis on a variable-by-variable basis (i.e., case-wise

omission). To be included in contingency tables, respondents needed to have provided data on all dependent and independent variables included in that table.

To reduce the number of contingency tables, dependent or independent variables with low numbers of responses were excluded. Variables to include as a priority in the secondary analysis were identified during the initial data summaries, discussed with my supervisors and documented in the statistical analysis plan. It was stipulated that independent variables might also be excluded from analysis if they failed to show any variance in responses - for example, if very few participants had a child take part in a study, or if all respondents reported that they are in good health.

5.9.2 Primary analyses- summaries and cross-tabulations

The first step in the analysis was a summary of questionnaire responses, broken down by source study, to show the pattern of responses to each question (or what participants preferences were). Response rate was calculated and a summary of missing data for each question is presented. The respondent characteristics were compared to their respective cohort profiles (where available) and to the general population.

5.9.3 Planned secondary analyses

The data summaries described above and presented in section 5.16 below provided insights into overall levels of concern regarding various data sharing scenarios, and preferences for procedures such as consent by presenting frequencies. The purpose of the secondary analyses was to assess whether there were any significant associations between the dependent and independent variables, and to identify whether independent variables may have exhibited any influence over attitudes and opinions. Taking gender as an example, do more (as a proportion) male or female respondents answer questions about data sharing in a certain way or were there no differences between genders?

For this I used cross-tabulations (or contingency tables) including measures of association (Pearson's Chi-squared statistic (p value)) with the addition of row and column percentages.

First, I produced cross-tabulations for the independent variables, with the Pearson's Chi squared statistic (p value) assessed. This identified any associations between the independent variables. All significant results from independent variable cross-tabulations can be viewed in Appendix J.

Then, each dependent variable was cross tabulated with each independent variable in turn, and those with a significant result were recorded (Appendix K). Table 53 (section 5.20.12, page 230) summarises the number of significant relationships identified with each independent variable broken down by the original research questions (1-3) of the study.

Then, for 'key' questionnaire questions (defined in section 5.9.4), the results of crosstabs with significant associations are discussed in terms of proportional results, for example did more males or females respond to a question in a certain way. As with the primary data summaries, this secondary research intends to address research questions 1-3 (see section 5.7 for a reminder of the research questions). As described in section 5.8.7 above, the responses to the dependent variables were dichotomised to render them suitable for analysis.

As part of the secondary analyses, I decided to run post-hoc analyses using the Bonferroni correction (McDonald, J, 2014) to examine each possible set of pairwise groupings of categories of the independent variable and thereby identify which of the group(s) led to a significant association overall. Bonferroni was selected as it was the method with which I was most familiar, and because it provided a middle ground between being too conservative and not conservative enough (UCLA, 2021). It involves setting the p-value at 0.05 divided by the number of paired comparisons; for example, in the case of deprivation score (5 levels) divided by 10. Appendix L provides a summary table of the number of contrasts and resultant new p-value. These post-hoc analyses were limited to comparisons between the dependent variables and independent variables that presented the highest number of significant associations for each of the 'key' variables described below.

5.9.4 Key variables

'Key' variables were identified as those that were critical to answering research questions 1-3 (attitudes towards sharing, whether knowing about sharing affects likelihood of taking part in studies and preferences for sharing). Key variables were also those that encompassed the themes explored in the systematic review and then subsequently in the grey literature (consent, storage, access and type of sharing). This maintains the link between the questionnaire results to those from the systematic review and the review of the grey literature allowing for triangulation of results later. Table 14 below identifies the key questionnaire questions.

	Research question	Key Sub questions	Questions used to answer this	Sub-items by question
RQ1.	Attitudes towards sharing	How concerned	Q5	-
			Q6	Q6_1 - Q6_9
			Q7	Q7_1 – Q7_8
		How likely to give permission	Q8	Q8_1 – Q8_6
RQ2.	Does knowing about data sharing affect the likelihood of respondents taking part in research?		Q15	-
RQ3.	What are respondent's preferences for sharing?	Consent	Q12	Q12_1 – Q12_5
			Q13	Q13_1 – Q13_6
		Storage/Access	Q19	-
			Q20	Q20_1 – Q20_6
		Ownership	Q21	Q21_1 – Q21_6
		Feedback	Q22	-

Table 14: Key variables for presentation of secondary results

5.9.5 Excluded variables:

Not all variables were selected for secondary analyses.

Question 1, asking who took part in a research study (you, your child, both, neither) was not answered by ALSPAC, and was therefore excluded from secondary analysis. The dependent variables regarding a child's participation in a study (Q2b & Q3b) were not included in secondary analyses as there were few answers to these questions (n=16 to each) and the decision had been made previously, and recorded in the analysis plan, not to include in analyses any variables with fewer than 200 responses. These questions had been included in the survey in anticipation of the FiCTION FUTURES participants forming one of the study sources.

Source study was presented in the frequency tables used to visually represent primary analysis and therefore was not presented in any presentation of the secondary analysis. Any significant associations between source study and the dependent variables can be viewed in Appendix K.

5.9.6 Significance:

Given the number of variables collected, the number of potential cross-tabulations for secondary analyses was high. As well as being time consuming, a large number of analyses could result in significant results being identified purely by chance. Results were therefore cautiously interpreted, considering the possible impacts of multiple testing. P values, odds ratios and 95% confidence intervals from the Stata output were recorded for reporting.

Only statistically significant results at a 0.05 significance level are reported below in the results (section 5.10) or the appendices (Appendix K) with corresponding 95% confidence intervals, grouped by each section of the questionnaire.

The data were also searched for non-significant results that were close to statistical significance (e.g., p=0.051) that might support those results presented below, but there were few results of this nature and those that were apparent were not focussed around one particular dependent or independent variable. Therefore, these non-significant results did not provide additional useful evidence regarding participant's attitudes and are not presented here.

5.10 RESULTS

This section first explores the survey response rate and the amount of missing data in the responses received. This is followed by a description of respondents' characteristics; how representative they are of the study population, and the results of cross tabulations of independent variables. Finally, the primary analysis; the questionnaire survey responses are presented, exploring respondents' views on data sharing i.e., the answers to research questions 1 to 3 for all respondents in the form of their answers to each of the dependent variables, or questionnaire questions e.g., 'would any of the following motivate you to share your data?'. Results are presented in order by questionnaire section (attitudes to sharing, consent, storage) in Tables 18 to 49.

These summary results are accompanied by a description of evidence from the secondary analysis; contingency tables or bar charts which detail which respondents (based on characteristics) were most likely to report which attitudes. This is accompanied by measures of association (Pearson's Chi-square analysis) to explain whether the relationship between the independent variable and dependent variable is statistically significant. A summary of responses to the free text question follows. The chapter is concluded with a summary of the pattern of significant results and key independent variables.

5.11 Response rate

A response rate has been calculated based upon the number of participants or estimated number of participants to whom the questionnaire survey was distributed (as reported above in section 5.6 Questionnaire distribution).

At the time of survey distribution, according to VOICE, there were approximately 3,000 members, and the SUPER group and SAIL contained 30 members between them. The questionnaires distributed to ALSPAC and ACONF participants were sent to a select number of participants based upon their communication preferences. Exact figures and calculated response rates are detailed below in Table 15.

	Total participants or total participants survey distributed to	**Responses received**	**% Response rate**
SAIL & SUPER group	30	12	40%
VOICE	3,000	51	1.7%
ACONF	1,400	395	28.2%
ALSPAC	5,858	1,226	20.9%
TOTAL	**10,288**	**1,684**	**16.4%**

Table 15- Estimated questionnaire response rate

The response rate was in excess of the 20% response rate anticipated in section 5.5, above for all but one group (SAIL and SUPER groups (40%), ACONF (28.2%) and ALSPAC (20.9%)). For Voice the response rate was approximately 1.7% which led to an overall response rate of 16.4%. One possible reason for the low response rate from VOICE could be lack of promotion. The questionnaire was only featured and therefore promoted via the VOICE weekly newsletter email when it first went live and not thereafter. Participants would have had to be browsing opportunities on the website to identify the survey in the weeks after it was first posted.

5.12 Patterns of missing data

A summary of missing data is reported in Table 15, below:

Question No.	Question	Number of missing responses	Percent
1	Have any of the following ever taken part in a health research study?	1,234	73.3
2a	Was YOUR participation as	1,421	84.4
2b	Was YOUR CHILD'S participation as	1,668	99.1
3a	What was the experience of taking part in a study like for YOU?	458	27.2
3b	What was the experience of taking part in a study like for YOUR CHILD?	1,668	99.1
5	How concerned would you be if you knew data from a study that you were involved in was being shared?	117	7.0
6	**How concerned would you be if you knew data was being shared with:**		
6_1	Researchers at the same organisation where your data was collected.	114	6.8
6_2	Researchers at a pharmaceutical company, e.g., for developing new medicines	118	7.0
6_3	Researchers at another university	116	6.9
6_4	Researchers at another hospital	121	7.2
6_5	Researchers in another country	116	6.9
6_6	A charity or not for profit organisation	116	6.9
6_7	The government	114	6.8
6_8	A student at a university	116	6.9
6_9	On the internet for anyone to use	117	7.0
7	**If data from a study in which you were involved was being shared, how concerned would you be about the following?**		
7_1	If could still be identified in the data	124	7.4
7_2	If my data could be used in research I don't approve of	127	7.5
7_3	If my data could be stolen	127	7.5
7_4	If my data could be used for making a profit e.g., advertising instead of research	126	7.5
7_5	If it would be embarrassing if my data was linked back to me	126	7.5
7_6	If people could misinterpret the data and come to the wrong conclusions	124	7.4
7_7	If the original research team didn't get credit for collecting the data	126	7.5
7_8	If it stopped researchers doing their own original research	123	7.3
8	**How likely would you be to give permission for your data to be shared for the following reasons?**		
8_1	To do research in a University	129	7.7
8_2	To do research in a hospital	135	8.0
8_3	To help a pharmaceutical company do research	132	7.8
8_4	To help the government study health problems	135	8.0
8_5	To inform the public about a health issue.	133	7.9
8_6	To help students get data for projects	136	8.1

Question No.	Question	Number of missing responses	Percent
9	**Below is a list of potential benefits of data sharing. Which of these make you feel more positive about data sharing?**		
9_1	Researchers can check each other's results and conclusions, making science more open.	n/a	n/a
9_2	Rarer diseases and conditions can be studied more easily using combined data, without having to wait for more studies.	n/a	n/a
9_3	Researchers can get quicker answers to scientific questions using data already collected.	n/a	n/a
9_4	Researchers can get the most out of participant's contribution (data) to their studies.	n/a	n/a
9_5	I can contribute to more research that affects me or my family.	n/a	n/a
10	**Would any of the following motivate you to allow your data to be shared?**		
10_1	Assured anonymity of the data shared	138	8.2
10_2	Understanding exactly how the data will be used	140	8.3
10_3	Knowing exactly who will access the data	142	8.3
10_4	Chance to understand my own condition better	147	8.7
10_5	Chance to help others by contributing to research	133	7.9
11	**Imagine that the researcher from the study you took part in wants to share your data with other researchers. How willing would you be for them to share anonymised details of your:**		
11_1	Age	131	7.8
11_2	Gender	135	8.0
11_3	Education	137	8.1
11_4	Employment	135	8.0
11_5	Height & weight	132	7.8
11_6	Mental health	137	8.1
11_7	Cancers	137	8.1
11_8	HIV infection	147	8.7
11_9	Other diseases or conditions	135	8.0
11_10	Family history of disease	133	7.9
11_11	Reproductive health	138	8.2
11_12	Medications being taken	133	7.9
11_13	Smoking behaviour	139	8.3
11_14	Alcohol use	138	8.2
11_15	Illegal drug use	142	8.4
12	How and when would you like to be asked to share your data?	174	10.3

Question No.	Question	Number of missing responses	Percent
13	**What information would you like to see on the consent form before you agree to share your data?**		
13_1	Explain that my data may be shared.	173	10.3
13_2	HOW the researchers will protect (anonymise) my identity.	177	10.5
13_3	Explanation of WHO might benefit from using my data	180	10.7
13_4	Details of WHERE the data will be stored.	183	10.9
13_5	Details of HOW the data will be stored.	188	11.2
13_6	Details of WHO the data might be shared with.	178	10.57
13a	None of the above would convince me to share my data	n/a	n/a
14	How important is it that you are informed on the consent form that your study data might be shared?	172	10.2
15	If you knew your data might be shared, what effect would it have on you taking part in a study?	171	10.2
16	Would you prefer to give consent separately for each type of organisation your data could be shared with?	174	10.3
17	Do you think a register of participants willing to share their study data is a good idea?	177	10.5
18	If a register of participants who are willing to share their study data existed, would you be willing to be named on it?	178	10.6
19	How would you prefer your study data to be stored?	186	11.1
20	If data has controlled access: Who do you think should give permission for data to be shared and used again?	189	11.2
22	Who do you think should 'own' the data collected during a study?	187	11.1
23	What is your gender?	188	11.2
24	Which age group do you belong to?	187	11.1
25	How would you describe your ethnicity?	189	11.2
26	What is your highest level of educational achievement?	188	11.2
27	How would you describe your overall health at the moment?	187	11.1
28	What is your postcode?	1,382	82.1
29	Do you have any further comments about data sharing or about this survey?	1,488	88.4

Table 16 Missing data summary by question

Question 13a does not have any missing values, as it was a check box question that respondents either agreed with or did not. The same is true for questions 9_1 to 9_5 and questions 21_1 to 21_6, responses were either checked (agreed with) or not checked.

Eleven percent of respondents overall declined to give their gender, age, ethnicity, education and overall health at the time of questionnaire completion. Many of these missing answers were from the same respondents i.e., all demographic data was missing for roughly one in ten respondents.

Questions 1a-2b (who took part, was that as a healthy volunteer or participant with a health condition) and 3b (child's experience of taking part) were not included as part of the survey administered to ALSPAC participants as it was known that they were taking part in a longitudinal study and that their participation was as part of a birth cohort. The answers to these questions for ALSPAC participants were therefore missing by construction. All respondents were asked questions 3a, about their experience of taking part, and then questions 5 to 27. Question 4 was removed prior to survey distribution but the questionnaire questions were not renumbered once the survey had been built in Qualtrics as the question codes and numbering were not displayed to participants.

The questions about data sharing (Q5-Q8) had approximately 6-8% of missing responses for each item. Question 9 was a list of statements to select so it is not clear whether respondents accidentally skipped a particular response or did not select it because it did not reflect their opinion. Question 10 had between 7 and 8% of responses missing for each option and question 11 had between 7 and 9% of answers missing. At question 12, ten percent of respondents failed to choose a response.

Questions thirteen to 16 (excluding question 15) concerned consent and had approximately 10-11% of data missing for each sub-question. By Questions 19, 20 and 22, the number of missing responses had increased slightly to 11%. Again, question 21, is a list of statements so it was not possible to distinguish between missing data and disagreement.

The slight trend toward higher rates of item non-response as respondents progressed through the questionnaire is perhaps indicative of questionnaire fatigue.

5.13 Sample characteristics

A total of 1,684 completed surveys were received from 3 different groups of respondents (n=1,226 from ALSPAC, 395 from ACONF, and 63 from PPI groups).

The majority of respondents were female (n=953, 63.8%) and were aged 25-44 (n=1,116, 74.7%), due to the large proportion (n=1,226, 72.8%) of respondents from the ALSPAC study. Two hundred and eighty-three (19.0%) respondents were aged 65-74. Most respondents were white (n=1,445, 97%) with few from Black, Asian, Chinese, mixed, or other ethnic groups. Most respondents were educated to degree level (557, 37.5%) followed by AS/A Levels (n=272, 18.3%) and other professional qualifications (n=237, 16%). The majority of respondents self-reported their health at the time of survey completion as 'good' (n=750, 50.1%) or 'excellent' (n=412, 27.5%). Only 71 respondents (4.8 %) reported that they had 'poor' or 'very poor' health.

Excluding missing values, 51.7% (n=153) of respondent's postcodes were categorised as belonging to quintile 5 (most affluent) and only 9.1% (n=27) were classified as quintile 1 (most deprived). The majority of respondents (n=145, 57.8%) from ACONF were classified by postcode as least deprived, falling into Carstairs quintile 5. The majority (n=18, 28.6%) of respondents from the PPI groups declined to give their postcode.

A summary of respondent characteristics is presented in Table 17 Respondent characteristics below.

Question	Response	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
Study Source	Source	1,684	100	395	23.5	1,226	72.8	63	3.7
Q1 Have any of the following ever taken part in a research study?	You	264	58.7	218	56.3	n/a*	n/a	46	73.0
	Your child	6	1.3	6	1.6	n/a*	n/a	0	0
	You and your child	12	2.7	9	2.3	n/a*	n/a	3	4.8
	Neither	103	22.9	92	23.8	n/a*	n/a	11	17.5
	Not sure	65	14.4	62	16.0	n/a*	n/a	3	4.8
	Total	450	100	387	100	n/a*	n/a	63	100
Q2a Was your participation in the study as:	A person who had the health condition being studied	56	21.3	38	17.7	n/a*	n/a	18	37.5
	A healthy volunteer	161	61.2	135	62.8	n/a*	n/a	26	54.2
	A person who is at risk of developing the condition being studied	8	3.0	5	2.3	n/a*	n/a	3	6.3
	Not sure	38	14.5	37	17.2	n/a*	n/a	1	2.1
	Total	263	100	215	100	n/a*	n/a	48	100
Q2b Was your child's participation in the study as	A child who had the health condition being studied	4	25.0	3	23.1	n/a*	n/a	1	33.3
	A healthy volunteer	12	75.0	10	76.9	n/a*	n/a	2	66.7
	A person who is at risk of developing the condition being studied	0	0	0	0	n/a*	n/a	0	0
	Not sure	0	0	0	0	n/a*	n/a	0	0
	Total	16	100	13	100	n/a*	n/a	3	100
	Very positive	639	42.9	66	30.8	550	44.9	23	46.9

Question	Response	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
Q3a What was the experience of taking part in a study like for you	Positive	614	41.2	79	36.9	517	42.2	18	36.7
	Neither positive or negative	215	14.4	61	28.5	147	12.0	7	14.3
	Negative	5	0.3	2	0.9	1	0.1	1	2.0
	Very negative	0	0	0	0	1	0.1	0	0
	Not applicable	6	0.4	3	1.4	3	0.2	0	0
	Not sure	10	0.7	3	1.4	7	0.6	0	0
	Total	1,489	100	214	100	1,226	100	49	100
Q3b What was the experience of taking part in a study like for your child	Very positive	4	25.0	3	23.1	n/a*	n/a	1	33.3
	Positive	5	31.3	3	23.1	n/a*	n/a	2	66.7
	Neither positive or negative	4	25.0	4	30.8	n/a*	n/a	0	0
	Negative	0	0	0	0	n/a*	n/a	0	0
	Very negative	0	0	0	0	n/a*	n/a	0	0
	Not applicable	0	0	0	0	n/a*	n/a	0	0
	Not sure	3	18.8	3	23.1	n/a*	n/a	0	0
	Total	16	100	13	100	n/a*	n/a	3	100
Q23 Gender	Male	529	35.4	170	52.0	338	30.5	21	34.4
	Female	953	63.8	157	48.0	757	68.3	39	63.9
	Other	13	0.9	0	0	13	1.2	1	1.6
	Total	1,495	100	327	100	1,108	100	61	100
Q24 Age group	18-24	1	0.1	0	0	0	0	0	0
	25-44	1,116	74.7	0	0	1,109	100	7	12.1
	45-64	84	5.6	64	19.6	0	0	20	34.5
	65-74	283	19.0	262	80.4	0	0	21	36.2
	75-84	9	0.6	0	0	0	0	10	17.2
	85 and over	1	0.1	0	0	0	0	0	0
	Total	1,493	100	326	100	1,109	100	58	100
Q25 Ethnicity	White	1,445	97.0	322	99.4	1,066	96.2	57	98.3
	Mixed/multiple ethnic group	30	2.0	0	0	30	2.7	0	0
	Asian/Asian British	6	0.4	0	0	0	0	0	0
	Black/African/Caribbean/Black British	0	0	0	0	0	0	0	0
	Chinese	0	0	0	0	0	0	0	0
	Other Ethnic group	9	0.6**	2**	0.6	12 **	1.1	1**	1.7
	Total	1,490	100	324	100	1,108	100	58	100
Q26 Highest educational achievement	No qualifications	36	2.4	21	6.6	14	1.3	1	1.6
	O Levels/CSE/GCSE or equivalent	131	8.8	48	15.1	79	7.1	4	6.6
	AS/A Levels or equivalent	272	18.3	51	16.0	216	19.5	5	8.2
	Degree (e.g., BA, BSc) or equivalent	557	37.5	69	21.7	469	42.3	19	31.2
	Higher degree (e.g., MSc, PhD) or equivalent	229	15.4	27	8.5	186	16.8	16	26.2
	Professional qualifications (e.g., nursing, accountancy, teaching)	237	16.0	91	28.6	132	11.9	14	23.0
	Other	24	1.6	11	3.5	13	1.2	2	3.3
	Total	1,486	100	318	100	1,109	100	61	100
Q27 Overall health at the moment	Excellent	412	27.5	67	20.5	340	30.7	5	8.3
	Good	750	50.1	167	51.1	553	49.9	30	50.0
	Average	257	17.2	69	21.1	169	15.2	19	31.7
	Poor	58	3.9	19	5.8	36	3.3	3	5.0
	Very poor	13	0.9	1	0.3	9	0.8	3	5.0
	Not sure	6	0.4	4	1.2	2	0.2	0	0
	Total	1,496	100	327	100	1,109	100	60	100
Deprivation quintile	1	27	9.1	13	5.2	n/a*	n/a	14	31.1
	2	38	12.8	29	11.6	n/a*	n/a	9	20.0

Question	Response	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
	3	43	14.5	30	12.0	n/a*	n/a	13	28.9
	4	35	11.8	34	13.6	n/a*	n/a	1	2.2
	5	153	51.7	145	57.8	n/a*	n/a	8	17.8
	Total	296	100	251	100	n/a*	n/a	45	100

Table 17 Respondent characteristics

*ALSPAC were not asked about involvement in a study as it was known they were in a longitudinal study

**Small numbers are grouped for tabulation purposes

5.14 Sample representativeness

No data was provided by ALSPAC and ACONF on the demographic breakdown for the subset of respondents that were contacted on my behalf; however, cohort profiles for all participants taking part in the ALSPAC and ACONF studies are available (Batty *et al.*, 2004; Boyd *et al.*, 2013). To check whether questionnaire respondents matched the profile of their respective cohorts, the demographic response data of respondents from ACONF and ALSPAC were checked against their study cohort demographics where this data was available.

The questionnaire respondents from ACONF were broadly representative of the cohort profile (Batty *et al.*, 2004) in terms of gender split. At the time of the survey, ACONF participants would have been aged approximately 65 to 70, and the vast majority of questionnaire respondents were in the age group 65-74 (see Table 16, above). One observable difference is that of deprivation score. Based on postcode, the majority of ACONF respondents were in the most affluent quintile, though at birth 74% of traced individuals had fathers with a manual occupation and this figure was 67.9% in 1962. However, both of those measures were recorded nearly 70 years ago, and we have to allow for social mobility (which is actually described in the cohort profile by Batty *et al* (Batty *et al.*, 2004)) across the cohort as a whole whilst simultaneously recognising that potentially more affluent participants of ACONF were more likely to respond to this survey.

The ALSPAC Children of the 90s cohort profile (Boyd *et al.*, 2013) indicates that there should be an even split between male and female respondents (49% female). However, female questionnaire respondents were over-represented at 68.3%. This supports ALSPACs own findings that recent responders are more likely to be female (Boyd *et al.*, 2013). Unsurprisingly those who gave their age reported that they were in the 25-44 age group. The

ALSPAC cohort is 96% white (Boyd *et al.*, 2013), and correspondingly 96.2% of ALSPAC respondents chose this option. ALSPAC themselves recognise that respondents to their (recent) surveys have been more likely to be female, white, and less likely to be eligible for free school meals (therefore more affluent).

I also wanted to determine whether my sample of respondents resembled the general population. I used data from the Office of National Statistics UK mid-year estimates of 2020 (Office of National Statistics, 2020) for age, and the 2011 census data for gender, ethnicity and health (Office of National Statistics, 2011) and compared this to the respondent's survey data. Results are displayed in Figure 9 below.

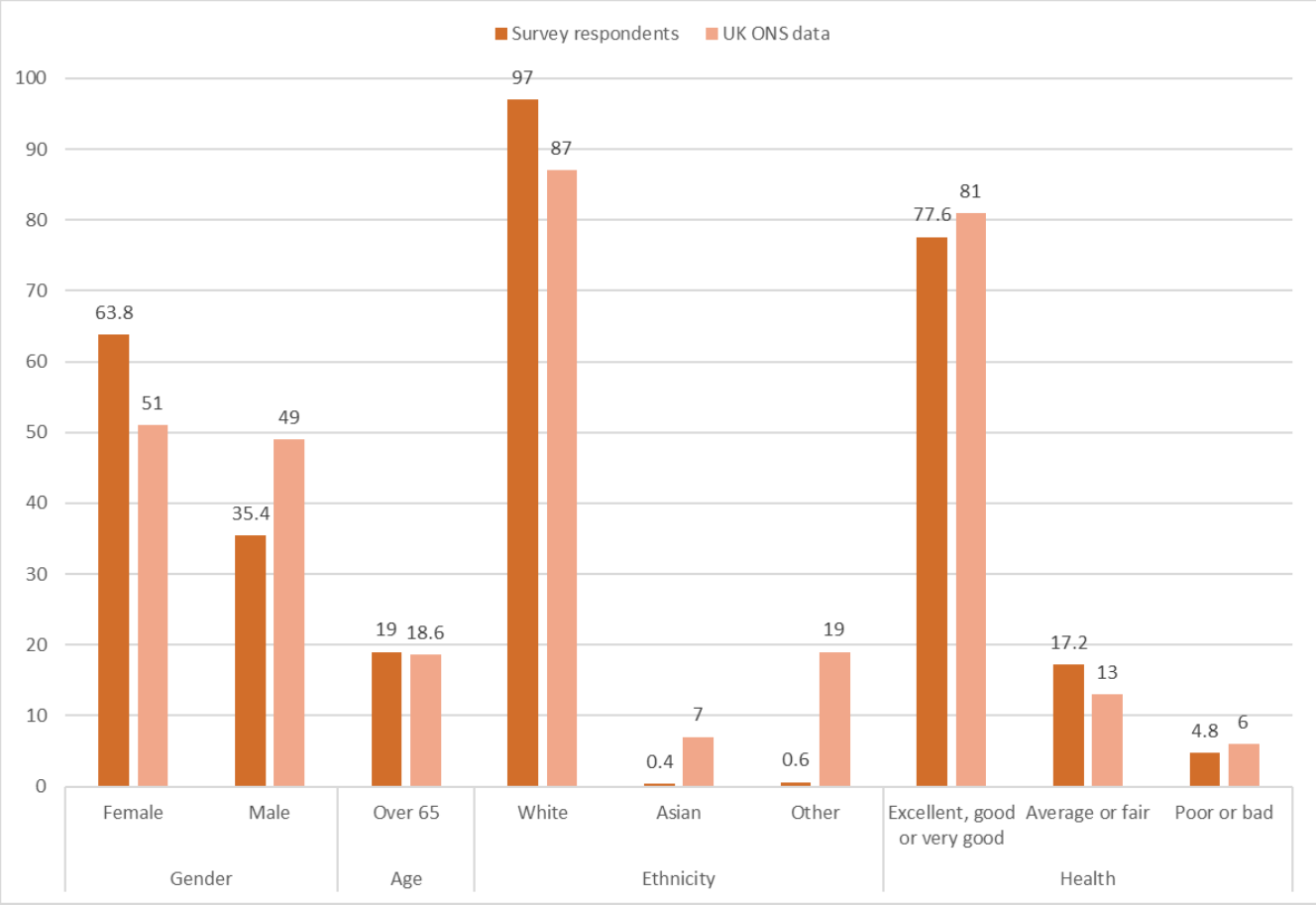


Figure 9 Respondent characteristics compared to those for the general population.

The Census data available from the ONS is from 2011, and so quite out of date, so the comparison made above is quite crude. However, we can see that the survey respondents were similar to the general population in some respects, such as proportion of the population over 65 and health, although the categories used to measure health were slightly different in each data set, and the questionnaire respondents tended to be either over 65 or

aged 25-44 because of the sampling frames used. The questionnaire respondents were less ethnically diverse than the general population and females were over-represented.

5.15 Independent variable crosstabs

Cross-tabulations were produced for all independent variables to determine whether there were any significant associations between the independent variables which could potentially influence results of the secondary analyses.

Age was significantly associated with gender ($p < 0.001$) and ethnicity ($p < 0.001$). Deprivation score was not significantly associated with health or education, but health and education were associated with each other ($p < 0.001$). Source (or the study in which respondents took part) and the experience of taking part were significantly associated with each other ($p < 0.001$).

The results that were identified as significant for all independent variables are tabulated in Appendix J for reference.

5.16 Questionnaire results:

The sections below cover each question from the questionnaire survey in turn and present the number and percentage of responses, and therefore the direction of majority opinion. Then, briefly, significant results from secondary analyses (cross-tabulations) for key variables (named in section 5.9.4) are presented, with a brief description of any corrections using the Bonferroni method. This is followed by further presentation of the cross-tabulations for variables that exhibited the highest number of significant associations with each other e.g., proportion of respondents 'concerned' by age.

5.17 Questions about taking part in research

Generally, respondents found taking part in a research study (Q3a) a 'positive' (n=614, 41.2%) or 'very positive' (n=639, 42.9%) experience. Of respondents (non-ALSPAC) who were asked questions 1, 2a, 2b, and 3b, the majority (n=264, 58.7%) reported that they themselves had taken part in a research study, and that they were a healthy volunteer (n=161, 61.2%). Only 1.3% (n=6) of non-ALSPAC respondents reported that their child had taken part in a study, 75% (n=12) of whom were taking part as a healthy volunteer.

5.18 Research Question 1: Attitudes towards sharing

5.18.1 Question 5: How concerned would you be if you knew data from a study that you were involved in was being shared?

The first question in the section of the questionnaire about attitudes towards data sharing (Q5) asked respondents how concerned they would be if they were informed that data from a study that they were involved in was being shared. The most common response was 'depends who it is shared with' (n=465, 29.7%) followed by 'not very concerned' (n=403, 25.7%). See Table 18 below.

Response:	Number of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Very concerned	94	6.0	22	6.2	67	5.8	5	8.1
Somewhat concerned	309	19.7	45	12.6	258	22.5	6	9.7
Not very concerned	403	25.7	105	29.3	286	24.9	12	19.3
Not at all concerned	279	17.8	90	25.1	167	14.6	22	35.5
Not sure	17	1.1	6	1.7	11	1.0	0	0
Depends who it is shared with	465	29.7	90	25.1	358	31.2	17	27.4
Total	1,567	100	358	100	1,147	100	62	100

Table 18: Responses to Question 5 How concerned would you be if you knew data from the study that you are involved in was being shared?

In secondary analyses, a Chi Square test of independence revealed that source study (χ^2 (d.f.=2, n=1,567) =28.3, p<0.001), age (χ^2 (d.f.=3, n=1,484) =33.8, p < 0.001) and respondents' self-rated health (χ^2 (d.f.=5, n=1487) =14.1, p0.015) were significantly associated with how 'concerned' respondents were regarding data sharing (compared to 'not concerned').

Post-hoc comparisons of age by 'concern' revealed significant differences in the proportion of respondents who were 'concerned' between respondents aged 25-44 and older age groups. Post-hoc comparisons of source study and 'concern' identified significant differences in the proportion of respondents 'concerned' between respondents from ACONF and ALSPAC. Further details are given in Table 19 below. All results from post-hoc analyses can be found in Appendix L.

Post-hoc comparisons between self-rated health and 'concern' revealed no significant differences between groups. This was inconsistent with the original analyses where a significant association was identified. Further post-hoc analysis was undertaken to examine this by combining categories with relatively small numbers; in this case 'poor' and 'very poor' were combined, while respondents who answered 'not sure' were excluded. No significant association was identified in this post-hoc analysis, so the inconsistency was resolved.

Question	Categories	p-value
How concerned would you be if you knew data from the study that you are involved in was being shared?	Age 25-44 vs 45-64	0.007
	Age 25-44 vs 65-74	<0.001
	ACONF vs ALSPAC	<0.001

Table 19: Q5 Bonferroni correction comparisons- age and self-rated health.

Proportionally, respondents from ALSPAC were more likely to be concerned (as compared to 'not concerned') about sharing than respondents from other groups with 60.1% of ALSPAC respondents reporting that they would be 'concerned' as compared to 45% of respondents from ACONF or the participant groups.

As displayed in Table 20, younger respondents (aged 25-44) were more likely to be concerned (as compared to 'not concerned') about sharing, with 60.4% of them reporting that they would be 'concerned' as compared to around 40% in the age groups 45-64 and 65-74 (Table 20).

	Number of respondents (%)					
Age	**Not Concerned**		**Concerned**		**Total**	
25-44	440	39.6	670	60.4	1,110	100
45-64	48	57.8	35	42.2	83	100
65-74	157	55.9	124	44.1	281	100
75-84	7	70	3	30	10	100
Total	652	43.9	832	56.1	1,484	100

Table 20: Responses to Question 5 by age

A higher proportion of respondents were 'concerned' as compared to 'not concerned' about their data being shared regardless of their health status (Table 21). Respondents who described their health as 'very poor' appeared to be much more concerned than respondents with other health statuses, with 84.6% respondents reporting that they would be 'concerned' as compared to between 52 and 61% for the other health groups. However, there were only 13 respondents with 'very poor' health. This is consistent with significant differences between groups being identified when post-hoc tests compared health status and 'concern' (Table 19).

	Number of respondents (%)					
Self-rated health	**Not Concerned**		**Concerned**		**Total**	
Excellent	192	46.8	218	53.2	410	100
Good	333	44.7	412	55.3	745	100
Average	98	38.3	158	61.7	256	100
Poor	27	47.4	30	52.6	57	100
Very poor	2	15.4	11	84.6	13	100
Not sure	0	0	6	100	6	100
Total	652	43.9	835	56.1	1,487	100

Table 21: Responses to Question 5 by self-rated health

5.18.2 Question 6: How concerned would you be if you knew your data was being shared with:

The majority of respondents were 'not at all concerned' about sharing with most organisations, although this unconcerned majority was less pronounced for sharing with a pharmaceutical company. More concern was exhibited for sharing with the government with 30.2% (n=474) of respondents 'somewhat concerned' and 18.6% (n=292) 'very concerned' than in respect of sharing with organisations such as universities, hospitals, or charities. Unsurprisingly, a spike in concern was observed for sharing data 'on the internet for anyone to use' with 61% (n=956) of respondents 'very concerned' about this. Full results for question 6 are given below in Table 22.

	Response:	**Number of respondents (%)**							
		All respondents		ACONF		ALSPAC		PPI groups	
Researchers at the same organisation where your data was collected	Very concerned	12	0.8	3	0.8	8	0.7	1	1.6
	Somewhat concerned	45	2.9	11	3.1	27	2.3	7	11.5
	Not very concerned	408	25.0	119	33.3	269	23.4	20	32.8
	Not at all concerned	1,094	69.7	219	61.3	842	73.1	33	54.1
	Not sure	11	0.7	5	1.4	6	0.5	0	0
	Total	1,570	100	357	100	1,152	100	61	100
Researchers at a pharmaceutical company, e.g., for developing new medicines	Very concerned	73	4.7	19	5.4	46	4.0	8	12.9
	Somewhat concerned	258	16.5	53	15.0	189	16.4	16	25.8
	Not very concerned	557	35.6	129	36.4	410	35.7	18	29.0
	Not at all concerned	652	41.6	146	41.2	487	42.4	19	30.7
	Not sure	26	1.7	7	2.0	18	1.6	1	1.6
	Total	1,566	100	354	100	1,150	100	62	100
	Very concerned	24	1.5	7	2.0	14	1.2	3	4.9

Researchers at a university	Somewhat concerned	116	7.4	27	7.6	79	6.9	10	16.4
	Not very concerned	582	37.1	141	39.7	427	37.1	14	23.0
	Not at all concerned	829	52.9	177	49.9	619	53.7	33	54.1
	Not sure	17	1.1	3	0.9	13	1.1	1	1.6
	Total	1,568	100	355	100	1,149	100	61	100
Researchers at a hospital	Very concerned	23	1.5	6	1.7	15	1.3	2	3.2
	Somewhat concerned	87	5.6	17	4.8	61	5.3	9	14.5
	Not very concerned	543	34.7	131	37.2	395	34.4	17	27.4
	Not at all concerned	899	57.5	196	55.7	670	58.3	33	53.2
	Not sure	11	0.7	2	0.6	8	0.7	1	1.6
	Total	1,563	100	352	100	1,149	100	62	100
Researchers in another country	Very concerned	193	12.3	61	17.0	121	10.5	11	17.7
	Somewhat concerned	405	25.8	92	25.7	297	25.9	16	25.8
	Not very concerned	441	28.1	97	27.1	328	28.6	16	25.8
	Not at all concerned	473	30.2	94	26.3	364	31.7	15	24.2
	Not sure	56	3.6	14	3.9	38	3.3	4	6.5
	Total	1,568	100	358	100	1,148	100	62	100
A charity or not for profit organisation	Very concerned	106	6.8	44	12.4	58	5.0	4	6.5
	Somewhat concerned	349	22.3	73	20.5	259	22.5	17	27.4
	Not very concerned	521	33.2	120	33.7	391	34.0	10	16.1
	Not at all concerned	539	34.4	99	27.8	413	35.9	27	43.6
	Not sure	53	3.4	20	5.6	29	2.5	4	6.5
	Total	1,568	100	356	100	1,150	100	62	100
The government	Very concerned	292	18.6	77	21.5	199	17.3	16	25.8
	Somewhat concerned	474	30.2	94	26.3	366	31.8	14	22.6
	Not very concerned	423	26.9	102	28.5	307	26.7	14	22.6
	Not at all concerned	326	20.8	66	18.4	243	21.1	17	27.4
	Not sure	55	3.5	19	5.3	35	3.0	1	1.6
	Total	1,570	100	358	100	1,150	100	62	100
A student at a university	Very concerned	217	13.8	54	15.1	150	13.1	13	21.0
	Somewhat concerned	388	24.7	70	19.6	305	26.5	13	21.0
	Not very concerned	497	31.7	137	38.4	346	30.1	14	22.6
	Not at all concerned	425	27.1	86	2.8	323	28.1	16	25.8
	Not sure	41	2.6	10	2.8	25	2.2	6	9.7
	Total	1,568	100	357	100	1,149	100	62	100
On the internet for anyone to use	Very concerned	956	61.0	240	67.6	681	59.2	35	56.5
	Somewhat concerned	367	23.4	63	17.8	286	24.9	18	29.0
	Not very concerned	111	7.1	20	5.6	89	7.7	2	3.2

Not at all concerned	95	6.1	21	5.9	70	6.1	4	6.5
Not sure	38	2.4	11	3.1	24	2.1	3	4.8
Total	1,567	100	355	100	1,150	100	62	100

Table 22: Responses Question 6 How concerned would you be if you knew data was being shared with:

In secondary analyses, Chi Square tests of independence revealed that all independent variables analysed except for age (i.e. respondents' gender, ethnicity, education, source study, education, deprivation quintile, experience of taking part and their rating of their overall health) were associated with respondents being 'concerned' (as compared to 'not concerned') about sharing with at least one of the nine potential organisations. The organisation(s) about which respondents were most concerned varied across these independent variables. Experience of taking part, self-rated health and source study were the independent variables which had the highest number of significant associations with being 'concerned'. There was no significant association between experience of taking part in research and questions 6a; sharing with researchers at the organisation where data was collected, and 6d; sharing with researchers at a hospital. All significant associations for Question 6 can be viewed in Appendix K.

Post-hoc comparisons of self-rated health and 'concern' about sharing data with various organisations generally revealed significant differences in proportions 'concerned' between respondents who rated their health as 'not sure' and more positive health statuses (excellent, good, average). Comparisons between experience of taking part (ETP) and 'concern' with sharing with pharmaceutical companies, identified significant differences in proportions 'concerned' between respondents who were 'not sure' about their experience and respondents who rated their experience as 'very positive'. Further details are given in Table 23 below. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
Researchers at a pharmaceutical company	Very positive (ETP) vs not sure	0.018
	Positive (ETP) vs not sure	0.036
Researchers at a university	Excellent vs not sure	<0.001
	Good vs not sure	<0.001
	Average vs not sure	<0.001
	Poor vs not sure	0.001
	Very poor vs not sure	0.006
Researchers at a hospital	Excellent vs not sure	0.001

	Good vs not sure	0.001
	Average vs not sure	0.002
	Poor vs not sure	0.006
Researchers in another country	Excellent vs very poor	0.017
	Good vs very poor	0.015
	Average vs very poor	0.046
A charity or not for profit organisation	Very positive (ETP) vs not sure	0.014
	Excellent vs not sure	0.044
	Excellent vs very poor	0.017
	Good vs very poor	0.040
A student at a university	Excellent vs not sure	0.026

Table 23: Q6 Bonferroni correction comparisons- self-rated health and experience of taking part.

Proportional results for the number of respondents concerned about sharing with various organisations by self-rated health and experience of taking part are displayed in Figures 10 and 11 below.

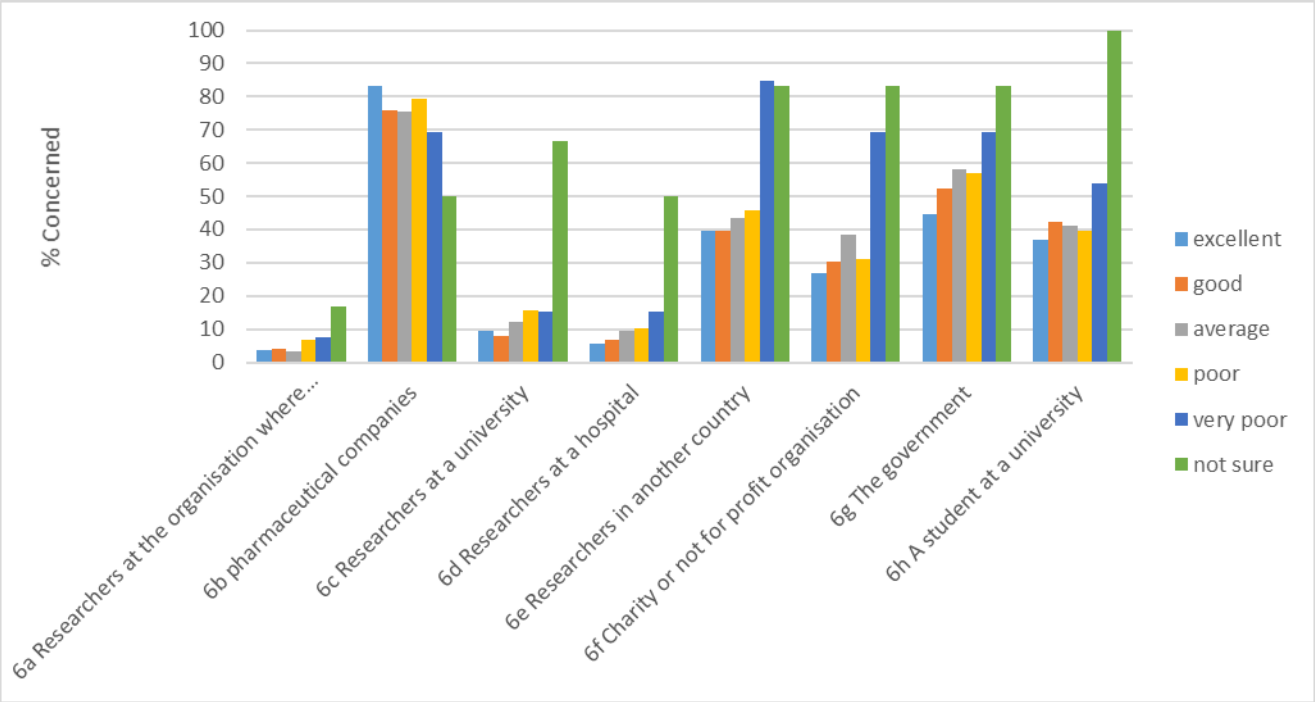


Figure 10: Responses to Question 6 by self-rated health

It is noticeable in Figure 10, that generally, respondents who self-rated their health as 'very poor' or 'not sure' and in some cases 'average' or 'poor' were more likely than respondents who rated their health as good or excellent to be concerned about sharing with most organisations. It is only when respondents were asked about sharing with a pharmaceutical

company that those who rated their health more positively were more or equally likely to exhibit concern. For question 6a, sharing data with researchers at the organisation where the data was collected, concern was lowest, and there was no significant association between this question and any of the independent variables.

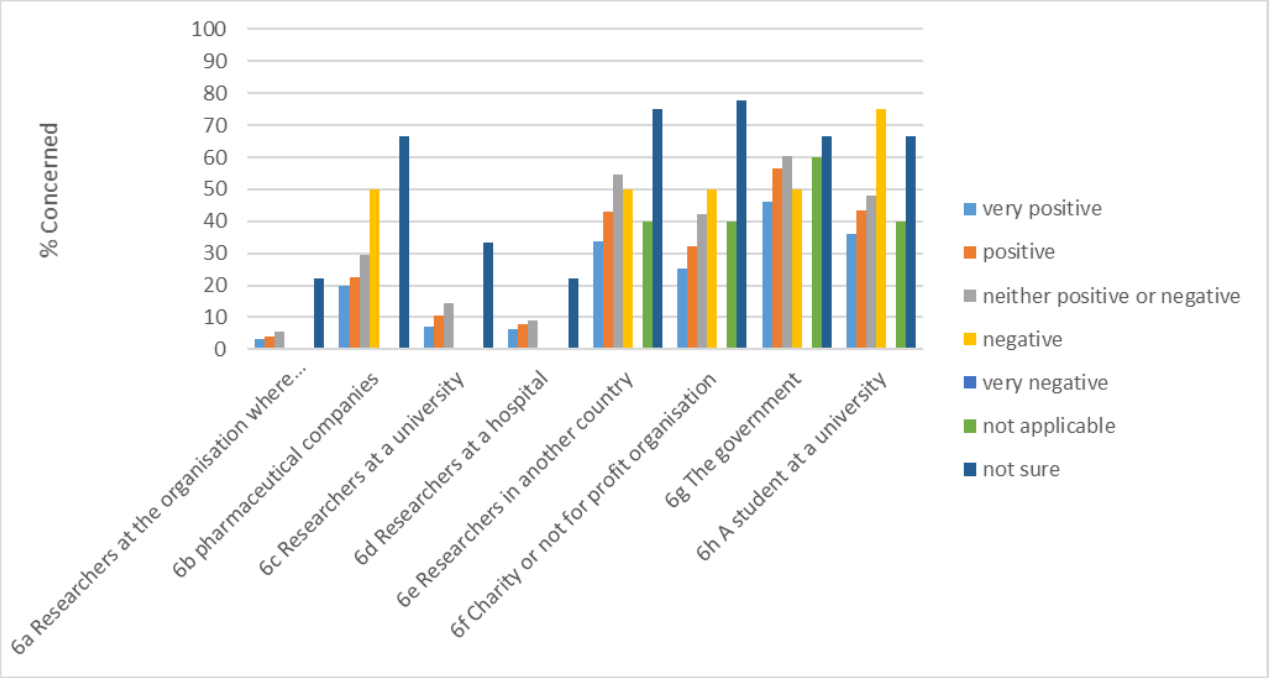


Figure 11: Responses to Question 6 by experience of taking part in research

Respondents who indicated that they were 'not sure' about their experience of taking part in research almost consistently exhibited the highest level of concern. It is not clear whether respondents selected 'not sure' to conceal whether or not they took part, or because they were unsure how to rate their experience (e.g.: because some aspects were positive and others negative). If we exclude respondents who answered 'not sure' or 'not applicable' Figure 11 shows that generally, the less positive the experience of taking part in research was, the more concerned respondents were about sharing with the various organisations. This pattern is not exhibited however, when asked about sharing with the government. Respondents who rated their experience as negative were also seen to be more 'concerned' than other respondents about sharing with a student.

5.18.3 Question 7: If data from a study in which you were involved was being shared, how concerned would you be about the following?

Respondents' main concerns regarding sharing were harms that can be categorised as security issues, such as being identified (n=1,009, 64.7%) or having their data stolen

(n=1,068, 68.6%). Respondents were more likely to be 'very concerned' about embarrassment if their data was linked back to them (n=850, 54.6%) than about data being used for profit (n=689, 44.2%), data being misinterpreted (n=691, 44.3%), or researcher-related issues such as lack of acknowledgement of the original research team (n=612, 39.3%). All results presented in Table 24 below:

	Response:	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
a) If I could still be identified in the data	Very concerned	1,009	64.7	203	51.4	763	66.4	43	70.5
	Somewhat concerned	408	26.2	112	28.4	286	24.9	10	16.4
	Not very concerned	96	6.2	1	6.6	65	5.7	5	8.2
	Not at all concerned	35	2.2	7	1.8	25	2.2	3	4.9
	Not sure	12	0.8	26	6.6	11	1.0	0	0
	Total	1,560	100	395	100	1,150	100	61	100
b) If my data could be used in research I don't approve of	Very concerned	664	42.7	187	47.3	438	38.2	39	62.9
	Somewhat concerned	554	35.6	103	26.1	438	38.2	13	21.0
	Not very concerned	195	12.5	9	2.3	154	13.4	7	11.3
	Not at all concerned	116	7.5	14	3.5	99	8.6	3	4.8
	Not sure	28	1.8	34	8.6	19	1.7	0	0
	Total	1,557	100	395	100	1,148	100	62	100
c) If my data could be stolen	Very concerned	1,068	68.6	282	81.5	742	64.6	44	71.0
	Somewhat concerned	356	22.9	46	13.3	296	25.8	14	22.6
	Not very concerned	90	5.8	2	0.6	74	6.4	4	6.5
	Not at all concerned	36	2.3	4	1.2	32	2.8	0	0
	Not sure	7	0.5	12	3.5	5	0.4	0	0
	Total	1,557	100	346	100	1,149	100	62	100
d) If my data could be used for making a profit e.g., advertising instead of research	Very concerned	689	44.2	0	0	689	60.0	0	0
	Somewhat concerned	631	40.5	267	77.0	322	28.0	42	67.7
	Not very concerned	172	11.0	60	17.3	96	8.4	16	25.8
	Not at all concerned	37	2.4	4	1.2	33	2.9	0	0
	Not sure	29	1.9	16	4.6	9	0.8	4	6.5
	Total	1,558	100	347	100	1,149	100	62	100
e) If it would be embarrassing if my data was linked back to me	Very concerned	850	54.6	212	61.1	597	52.0	41	66.1
	Somewhat concerned	409	26.3	95	27.4	303	26.4	11	17.7
	Not very concerned	179	11.5	23	6.6	148	12.9	8	12.9

	Response:	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
	Not at all concerned	100	6.4	11	3.2	87	7.6	2	3.2
	Not sure	20	1.3	6	1.7	14	1.2	0	0
	Total	1,558	100	347	100	1,149	100	62	100
f) If people could misinterpret the data and come to the wrong conclusions	Very concerned	691	44.3	215	62.0	433	37.6	43	69.4
	Somewhat concerned	541	24.7	109	31.4	418	36.3	14	22.6
	Not very concerned	221	14.2	15	4.3	203	17.6	3	4.8
	Not at all concerned	81	5.2	3	0.9	76	6.6	2	3.2
	Not sure	26	1.7	5	1.4	21	1.8	0	0
	Total	1,560	100	347	100	1,151	100	62	100
g) If the original research team didn't get credit for collecting the data	Very concerned	612	39.3	157	45.2	429	37.3	26	41.9
	Somewhat concerned	596	38.3	145	41.8	424	36.9	27	43.6
	Not very concerned	212	13.6	33	9.5	175	15.2	4	6.5
	Not at all concerned	102	6.6	8	2.3	90	7.8	4	6.5
	Not sure	36	2.3	4	1.2	31	2.7	0	0
	Total	1,558	100	347	100	1,149	100	62	100
h) If it stopped researchers doing their own original research	Very concerned	583	37.4	162	46.6	389	33.8	32	51.6
	Somewhat concerned	521	33.4	139	39.9	357	31.0	25	40.3
	Not very concerned	279	17.9	35	10.1	242	21.0	2	3.2
	Not at all concerned	108	6.9	7	2.0	100	8.7	1	1.6
	Not sure	70	4.5	5	1.4	63	5.5	2	3.2
	Total	1,561	100	348	100	1,151	100	62	100

Table 24: Responses to Question 7 If data from the study in which you were involved was being shared, how concerned would you be about the following?

In secondary analyses, Chi Square tests of independence revealed that all independent variables analysed (respondents' age, gender, ethnicity, education, source study, education, deprivation quintile, experience of taking part and their rating of their overall health) were associated with respondents being 'concerned' (as compared to 'not concerned') about at least one of the eight potential harms related to data sharing. Age, gender, and source study were the independent variables with the highest number of significant associations with harms. There was no significant association between age and question 7a 'if I could still be identified…' and no significant association between gender and questions 7a, 7c, 7f, and 7h. Details of all significant associations for Question 7 can be viewed in Appendix K.

Post-hoc comparisons of age by 'concern' about potential harms revealed significant differences in proportions 'concerned' between respondents aged 25-44 and respondents aged 65-74 for all harms that had a significant association with age. Significant differences in proportions 'concerned' were also observed between males and females for some harms. Further details are given in Table 25 below. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
If my data could be used in research I don't approve of	Male vs female	0.002
	Age 25-44 vs 65-74	0.023
If my data could be stolen	Age 25-44 vs 65-74	0.047
If my data could be used to make a profit	Age 25-44 vs 65-74	0.002
	Age 25-44 vs 75-84	0.002
	Age 65-74 vs 75-84	0.035
	Male vs female	0.001
If it would be embarrassing if my data was linked back to me	Age 25-44 vs 45-64	0.033
	Age 25-44 vs 65-74	0.001
	Male vs female	0.016
If people could misinterpret the data and come to the wrong conclusions	Age 25-44 vs 45-64	<0.001
	Age 25-44 vs 65-74	<0.001
If the original research team didn't get credit…	Age 25-44 vs 65-74	0.001
	Male vs female	<0.001
If it stopped researchers doing their own original research	Age 25-44 vs 45-64	<0.001
	Age 25-44 vs 65-74	<0.001

Table 25: Q7 Bonferroni correction comparisons- age and gender.

The potential harms around sharing about which respondents were 'concerned' varied across these independent variables. Results for the number of respondents concerned about various harms associated with sharing by age and gender are displayed below in Figures 12 and 13.

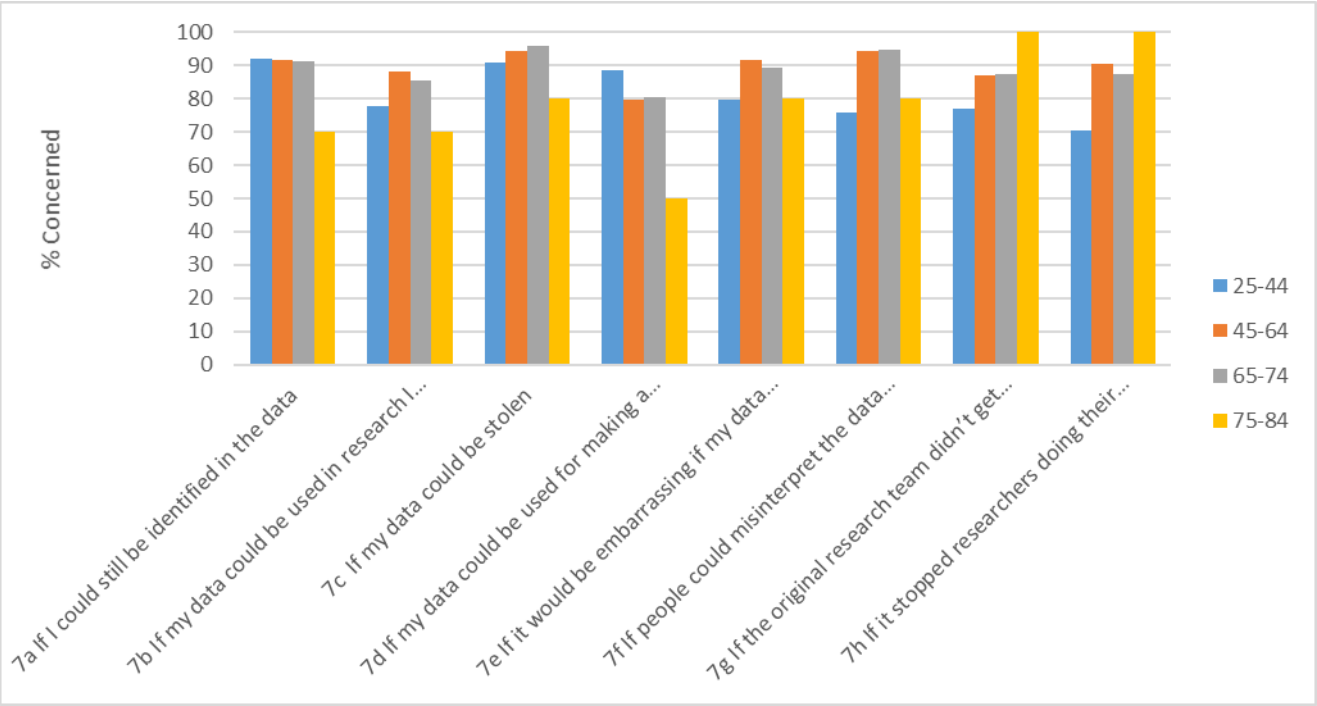


Figure 12: Responses to Question 7 by age

As exhibited in Figure 12, respondents aged 75-84 were less likely to be 'concerned' than younger age groups about the potential harms associated with sharing, that is until presented with researcher issues; questions 7g (if the original research team didn't get credit) and 7h (if it stopped researchers doing their own original research), where it appears that respondents aged 75-84 are more likely to be 'concerned' than younger age groups. Fewer respondents aged 75-84 (50%) appeared to be 'concerned' about their data being used to make a profit, than they were about other potential harms (70-100% concerned) and were also less likely to be 'concerned' about this than other age groups (79-88% concerned). Generally, there was little difference in proportions 'concerned' between respondents aged 45-64 and 65-74 for all potential harms.

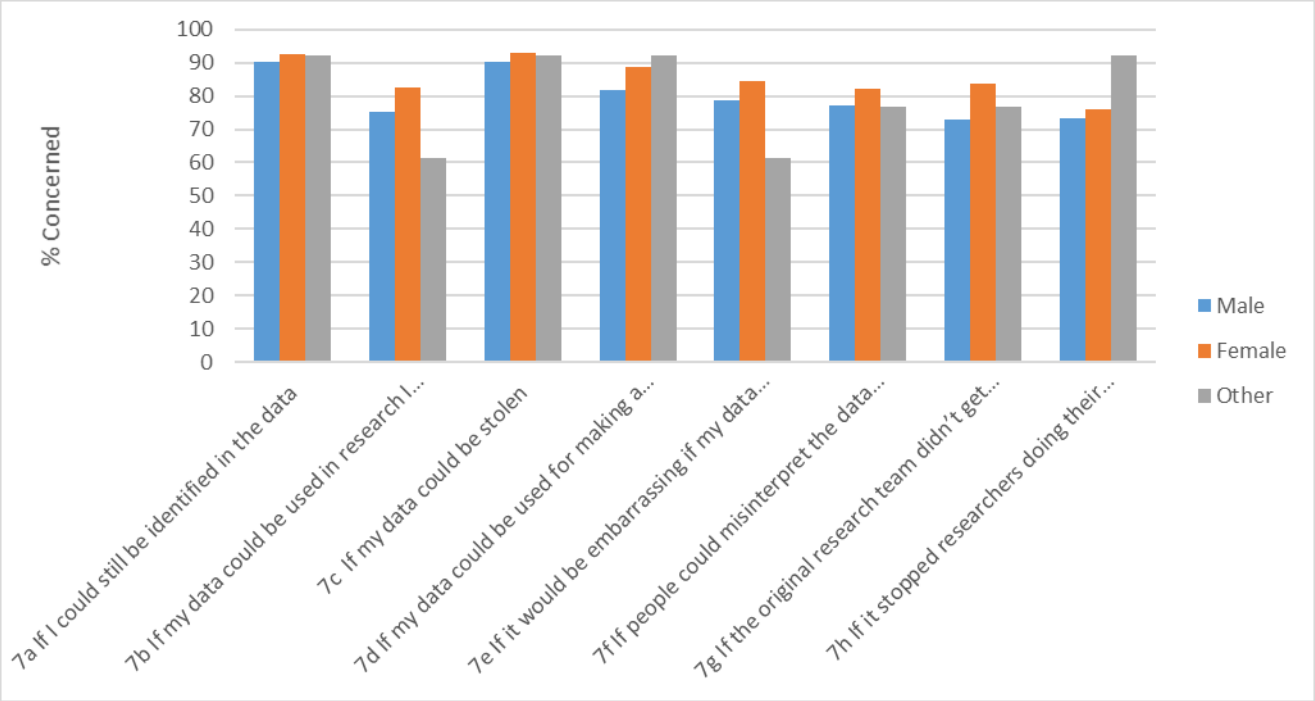


Figure 13: Responses to Question 7 by gender

As exhibited in Figure 13, female respondents were more likely to be 'concerned' about all potential harms from sharing than males, although only by between 2 and 11% more. Respondents who gave their gender as 'other' appear to be much more likely to be concerned about sharing preventing researchers from doing their own research (7h), with 92.3% of respondents giving an answer that indicated they would be 'concerned' as compared to 75% for males and females, but it must be noted that only 13 respondents actually gave their gender as 'other'.

5.18.4 Question 7a: Which of the above statements is of most concern to you?

Of all the potential harms of sharing, respondents who answered question 7a were most concerned about being identified in the data.

Response:	Number of respondents (%)							
	All participants		ACONF		ALSPAC		PPI Groups	
a) If I could still be identified in the data	715	46.1	110	32.5	580	50.4	25	40.3
b) If my data could be used in research I don't approve of	154	9.9	41	12.1	102	8.9	11	17.7
c) If my data could be stolen	325	21.0	93	27.4	226	19.7	6	9.7
d) If my data could be used for making a profit e.g., advertising instead of research	152	9.8	43	12.7	102	8.9	7	11.3
e) If it would be embarrassing if my data was linked back to me	84	5.4	14	4.1	67	5.8	3	4.8
f) If people could misinterpret the data and come to the wrong conclusions	48	3.1	19	5.6	25	2.2	4	6.5

g) If the original research team didn't get credit for collecting the data		23	1.5	2	0.6	21	1.8	0	0
h) If it stopped researchers doing their own original research		50	3.2	17	5.0	27	2.4	6	9.7
TOTAL		1,551	92.1	339	85.8	1,150	93.8	62	98.4

Table 26: Responses to Question 7a Which of the above statements is of MOST concern to you?

5.18.5 Question 8: How likely would you be to give permission for your data to be shared for the following reasons?

When asked about the likelihood of granting permission for their data to be used for various purposes, the majority of respondents indicated that they were 'very likely' to share for research in a university (n=974, 62.6%), a hospital (n=1,119, 72.2%) or to inform the public about a health issue (n=742, 47.8%). The likelihood of the majority of respondents agreeing to share with pharmaceutical companies (n=615, 39.6%), to help students get data for a project (n=578, 37.3%) and 'to help the government study health problems' (n=606, 39.1%) was less assured; given as 'somewhat'.

		Number of respondents (%)							
	Response:	**All respondents**		**ACONF**		**ALSPAC**		**PPI groups**	
To do research in a University	Very likely	974	62.6	231	67.2	702	61.0	41	67.2
	Somewhat likely	467	30.0	86	25.0	364	31.7	17	27.9
	Neither likely or unlikely	69	4.4	21	6.1	47	4.1	1	1.6
	Somewhat unlikely	27	1.7	3	0.9	23	2.0	1	1.6
	Very unlikely	18	1.2	3	0.9	14	1.2	1	1.6
	Total	1,555	100	344	100	1,150	100	61	100
To do research in a hospital	Very likely	1,119	72.2	266	77.6	811	70.8	42	68.9
	Somewhat likely	371	24.0	61	17.8	294	25.7	16	26.2
	Neither likely or unlikely	24	1.6	10	2.9	13	1.1	1	1.6
	Somewhat unlikely	19	1.2	3	0.9	16	1.4	0	0
	Very unlikely	16	1.0	3	0.9	11	1.0	2	3.3
	Total	1,549	100	343	100	1,145	100	61	100
To help a pharmaceutical company do research	Very likely	559	36.0	122	35.6	423	36.9	14	23.0
	Somewhat likely	615	39.6	129	37.6	459	40.0	27	44.3
	Neither likely or unlikely	205	13.2	53	15.5	142	12.4	10	16.4
	Somewhat unlikely	122	7.9	31	9.0	86	7.5	5	8.2
	Very unlikely	51	3.3	8	2.3	38	3.3	5	8.2
	Total	1,552	100	343	100	1,148	100	61	100
To help the government study health problems	Very likely	552	35.6	121	35.5	406	35.4	25	41.0
	Somewhat likely	606	39.1	134	39.3	450	39.2	22	36.1
	Neither likely or unlikely	201	13.0	50	14.7	145	12.6	6	9.8

	Response:	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
	Somewhat unlikely	126	8.1	22	6.5	99	8.6	5	8.2
	Very unlikely	64	4.1	14	4.1	47	4.1	3	4.9
	Total	1,549	100	341	100	1,147	100	61	100
To inform the public about a health issue.	Very likely	742	47.8	171	50.2	537	46.7	34	55.7
	Somewhat likely	554	35.7	113	33.1	424	36.9	17	27.9
	Neither likely or unlikely	170	11.0	41	12.0	123	10.7	6	9.8
	Somewhat unlikely	59	3.8	13	3.8	44	3.8	2	3.3
	Very unlikely	26	1.7	3	0.9	21	1.8	2	3.3
	Total	1,551	100	341	100	1,149	100	61	100
To help students get data for projects	Very likely	491	31.7	110	32.5	355	30.9	26	42.6
	Somewhat likely	578	37.3	145	42.5	414	36.1	19	31.2
	Neither likely or unlikely	294	19.0	56	16.5	229	20.0	9	14.8
	Somewhat unlikely	134	8.7	19	5.6	109	9.5	6	9.8
	Very unlikely	51	3.3	9	2.7	41	3.6	1	1.6
	Total	1,548	100	339	100	1,148	100	61	100

Table 27: Responses to Question 8 How likely would you be to give permission for your data to be shared for the following reasons?

In secondary analyses, Chi squared tests of independence revealed that all independent variables analysed (respondents' age, gender, ethnicity, education, source study, deprivation quintile, education, experience of taking part and their rating of their overall health) were significantly associated with likelihood ('likely' versus 'not likely') of respondents granting permission for their data to be used for at least one of the various purposes put forward.

The number of significant associations between likelihood of respondents agreeing to share their data for various purposes varied across these independent variables. Self-rated health and experience of taking part in research were the independent variables with the highest number of significant associations with proposed sharing reasons. All significant associations for Question 8 can be viewed in Appendix K.

Post-hoc comparisons of experience of taking part by 'likelihood' identified significant differences between respondents who had a 'very positive' and/or 'positive' experience and those who stated that they had a 'neither positive or negative' experience. Further details are given in Table 28 below. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
To do research in a university	Very positive vs neither	<0.001
	Positive vs neither	0.010
To help a pharmaceutical company do research	Very positive vs neither	0.027
To help the government study health problems	Very positive vs neither	0.000
To inform the public about a health issue	Very positive vs neither	<0.001
	Positive vs neither	0.044
To help students get data for projects	Very positive vs neither	<0.001

Table 28: Q8 Bonferroni correction comparisons- self rated health and experience of taking part in research.

Proportional results for the number of respondents likely to share data for the various purposes suggested by health rating and experience of taking part are displayed below in Figures 14 and 15.

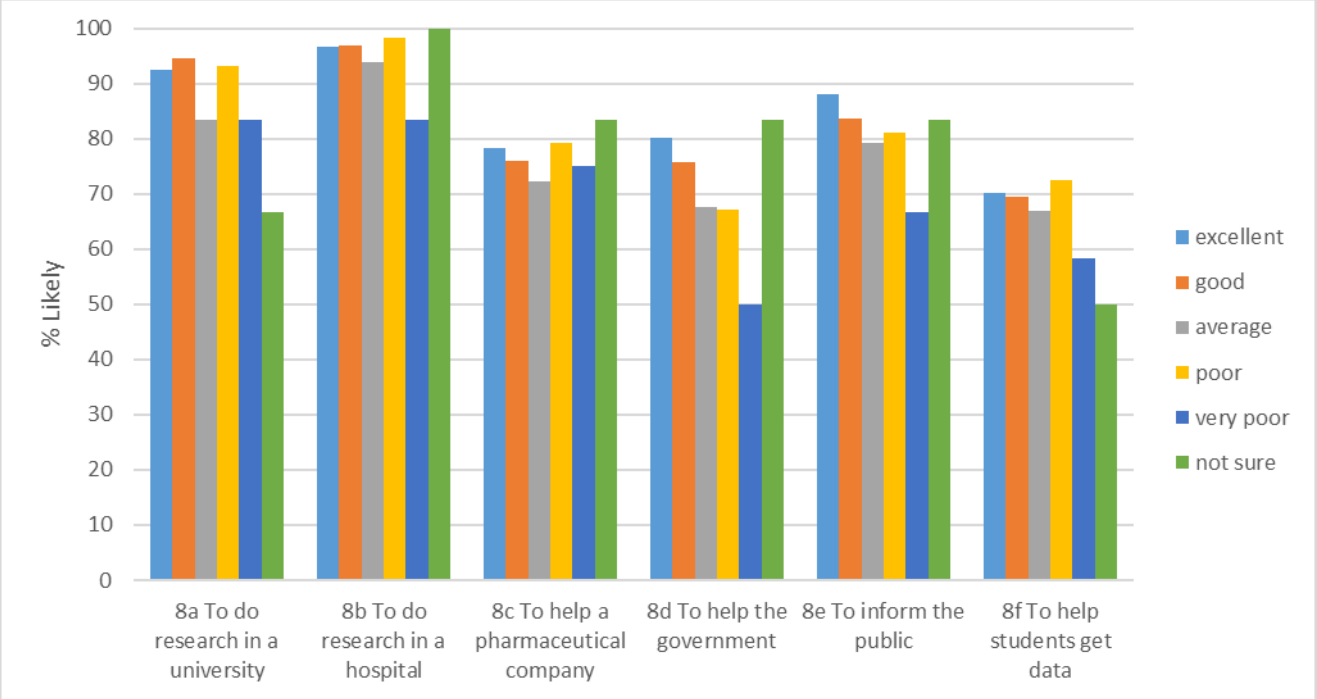


Figure 14: Responses to Question 8 by health rating

Figure 14 demonstrates that there is not a great deal of difference between respondents in terms of their self-rated health and their likelihood of sharing with various organisations. On closer inspection we can see that respondents who rated their health as 'not sure' were more likely to share their data for research in a hospital, with a pharmaceutical company and with the government than respondents with any other health rating. If we ignore those who give 'not sure' as their answer, there is a general pattern that the better a respondent rates

their health, the more likely they are to share for research in a university or hospital, help the government, inform the public, or share to help students.

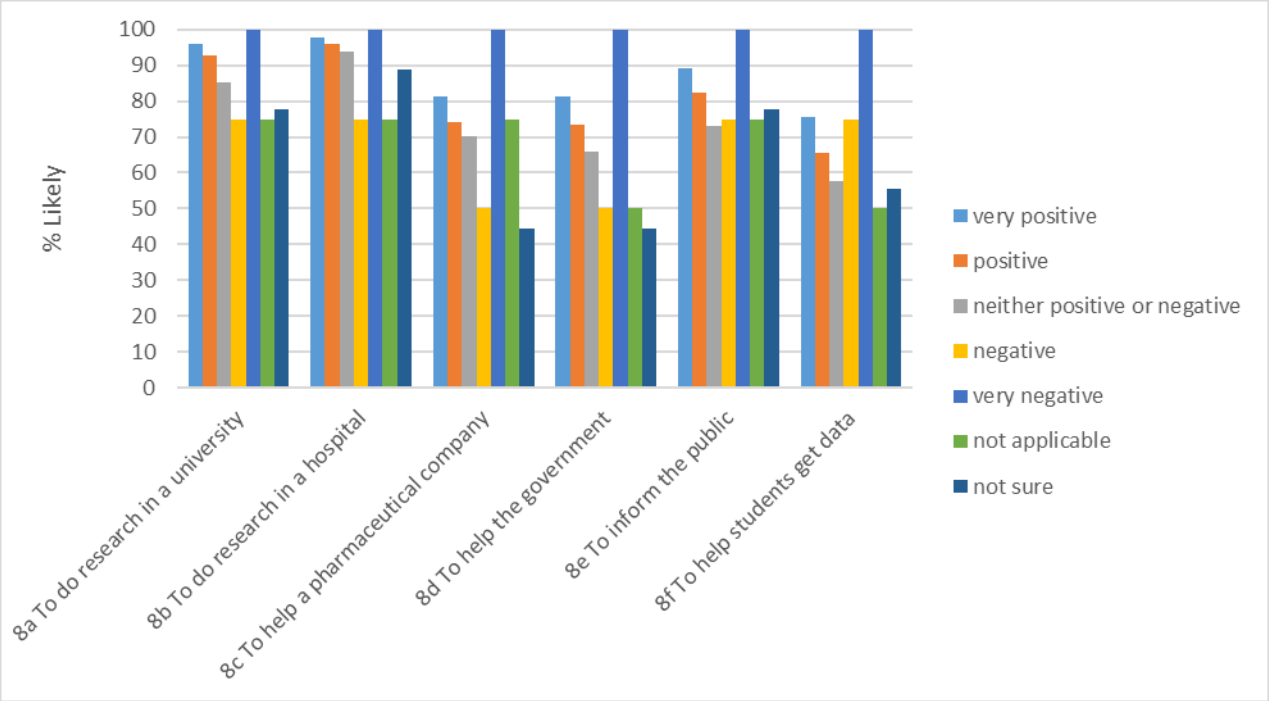


Figure 15: Responses to Question 8 by experience of taking part

As exhibited in Figure 15, generally, the more positively a respondent rated their experience of taking part, the more likely they were to agree to share their data with all recipients. The exception to this is the respondents who rated their experience as 'very negative', although there were comparatively few respondents who rated this way (0.9%), compared to those who rated their experience as good (50.1%) or very good (27.5%).

5.18.6 Question 9: Below is a list of potential benefits of sharing. Which of these make you feel more positive about data sharing?

Question 9 asked respondents which statements of potential benefit made them feel more positive about data sharing. All statements were popular, with approximately 60-80% of respondents selecting each of them, but the aspect of sharing that respondents found most beneficial, was 'Rarer diseases and conditions can be studied more easily using combined data, without having to wait for more studies.

Response:	Number of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Researchers can check each other's results and conclusions, making science more open.	1,149	68.2	274	69.4	863	70.4	51	81.0
Rarer diseases and conditions can be studied more easily using combined data, without having to wait for more studies.	1,374	81.6	313	79.2	1,051	85.7	53	84.1
Researchers can get quicker answers to scientific questions using data already collected.	1,197	71.1	286	72.5	900	73.4	52	82.5
Researchers can get the most out of participant's contribution (data) to their studies.	1,006	59.7	243	61.5	752	61.3	52	82.5
I can contribute to more research that affects me or my family.	1,029	61.1	236	59.8	787	64.2	57	90.5

Table 29: Responses to Question 9 Below is a list of potential benefits of data sharing. Which of these make you feel more positive about data sharing?

5.18.7 Question 10: Would any of the following motivate you to allow your data to be shared?

In terms of motivations to share (Q10), again, all of the potential motivations were popular with respondents, but privacy (assured anonymity) and altruism (chance to help others) were equally the most important to respondents.

	Response:	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
Assured anonymity of the data shared	Yes	1,442	93.3	308	92.2	1,078	93.7	56	91.8
	No	30	1.9	13	3.9	16	1.4	1	1.6
	Not sure	74	4.8	13	3.9	57	5.0	4	6.6
	Total	1,546	100	334	100	1,151	100	61	100
Understanding exactly how the data will be used	Yes	1,340	86.8	278	83.2	1,012	88.0	50	83.3
	No	102	6.6	26	7.8	71	6.2	5	8.3
	Not sure	102	6.6	30	9.0	67	5.8	5	8.3
	Total	1,544	100	334	100	1,150	100	60	100
Knowing exactly who will access the data	Yes	1,300	84.3	285	85.3	966	84.2	49	81.7
	No	112	7.3	22	6.6	82	7.1	8	13.3
	Not sure	130	8.4	27	8.1	100	8.7	3	5.0
	Total	1,542	100	334	100	1,148	100	60	100
Chance to understand my own condition better	Yes	1,050	68.3	264	80.5	739	64.3	47	78.3
	No	60	3.9	30	9.2	24	2.1	6	10.0
	Not sure	93	6.1	34	10.4	52	4.5	7	11.7
	Not Applicable	334	21.7	-	-	334	29.1	-	-
	Total	1,537	100	328	100	1,149	100	60	100
Chance to help others by contributing to research	Yes	1,441	92.9	326	95.88	1.058	92.0	57	93.44
	No	27	1.7	2	0.59	24	2.09	1	1.64
	Not sure	83	5.4	12	3.53	68	5.91	3	4.92
	Total	1,551	100	340	100	1,150	100	61	100

Table 30: Responses to Question 10 Would any of the following motivate you to allow your data to be shared?

5.18.8 Question 11: Imagine that the researcher from the study you took part in wants to share your data with other researchers. How willing would you be for them to share anonymised details of your:

Question 11 asked respondents which details about themselves (such as age, employment, alcohol use) they would be willing to share in an anonymised data set. The majority of respondents (between 45 and 58 percent) were 'very willing' to share all fifteen potential details. Details of mental health (45.6%) and employment (45.4%) had the lowest amounts of respondents who were very willing, but only marginally as compared to family history of disease (47.9%). Perhaps respondents value privacy for their family just as much as they do for themselves. Few respondents were 'not at all willing' to share any of the fifteen types of data.

	Response:	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
Age	Not at all willing	11	0.7	2	0.6	8	0.7	1	1.6
	Not very willing	17	1.1	5	1.5	10	0.9	2	3.2
	Not sure	42	2.7	12	3.5	28	2.4	2	3.2
	Willing	595	38.3	133	39.1	444	38.6	18	29.0
	Very willing	888	57.2	188	55.3	661	57.4	39	62.6
	Total	1,553	100	340	100	1,151	100	62	100
Gender	Not at all willing	12	0.8	1	0.3	10	0.9	1	1.6
	Not very willing	12	0.8	3	0.9	9	0.8	0	0
	Not sure	37	2.4	8	2.4	27	2.4	2	3.2
	Willing	590	38.1	137	40.5	434	37.8	19	30.7
	Very willing	898	58.0	189	55.9	669	58.2	40	64.5
	Total	1,549	100	338	100	1,149	100	62	100
Education	Not at all willing	27	1.8	3	0.9	22	1.9	2	3.3
	Not very willing	47	3.0	4	1.2	42	3.7	1	1.6
	Not sure	115	7.4	21	6.2	89	7.8	5	8.2
	Willing	583	37.7	135	39.9	431	37.5	17	27.9
	Very willing	775	50.1	175	51.8	564	49.1	36	59.0
	Total	1,547	100	338	100	1,148	100	61	100
Employment	Not at all willing	59	3.8	4	1.2	53	4.6	2	3.2
	Not very willing	82	5.3	7	2.1	72	6.3	3	4.8
	Not sure	186	12.0	22	6.5	158	13.8	6	9.7
	Willing	519	33.5	137	40.4	363	31.6	19	30.7
	Very willing	703	45.4	169	49.9	502	43.7	32	51.6
	Total	1,549	100	339	100	1,148	100	62	100
Height & weight	Not at all willing	20	1.3	4	1.2	15	1.3	1	1.6
	Not very willing	34	2.2	7	2.1	25	2.2	2	3.2
	Not sure	97	6.3	10	2.9	78	6.8	9	14.5
	Willing	601	38.7	143	42.1	443	38.5	15	24.2
	Very willing	800	51.6	176	51.8	589	51.2	35	56.5
	Total	1,552	100	340	100	1,150	100	62	100
Mental health	Not at all willing	35	2.3	3	0.9	31	2.7	1	1.6
	Not very willing	74	4.8	9	2.7	61	5.3	4	6.5
	Not sure	188	12.2	30	8.9	150	13.1	8	12.9
	Willing	545	35.2	136	40.2	393	34.3	16	25.8
	Very willing	705	45.6	160	47.3	512	44.6	33	53.2
	Total	1,547	100	338	100	1,147	100	62	100
Cancers	Not at all willing	24	1.6	1	0.3	22	1.9	1	1.6
	Not very willing	31	2.0	5	1.5	24	2.1	2	3.3
	Not sure	145	9.4	27	8.0	113	9.8	5	8.2
	Willing	555	35.9	129	38.2	409	35.6	17	27.9
	Very willing	792	51.2	176	52.1	580	50.5	36	59.0
	Total	1,547	100	338	100	1,148	100	61	100
HIV infection	Not at all willing	50	3.3	10	3.0	38	3.3	2	3.2
	Not very willing	47	3.1	6	1.8	38	3.3	3	4.8
	Not sure	209	13.6	52	15.5	150	13.2	7	11.3
	Willing	505	32.9	118	35.2	371	32.5	16	25.8
	Very willing	726	47.2	149	44.5	543	47.6	34	54.8
	Total	1,537	100	335	100	1,140	100	62	100
Other diseases or conditions	Not at all willing	22	1.4	1	0.3	20	1.7	1	1.6
	Not very willing	29	1.9	6	1.8	21	1.8	2	3.2
	Not sure	177	11.4	38	11.3	132	11.5	7	11.3
	Willing	568	36.7	130	38.6	423	36.8	15	24.2
	Very willing	753	48.6	162	48.1	554	48.2	37	59.7
	Total	1,549	100	337	100	1,150	100	62	100
Family history of disease	Not at all willing	26	1.7	3	0.9	22	1.9	1	1.6
	Not very willing	40	2.6	4	1.2	35	3.0	1	1.6
	Not sure	150	9.7	18	5.3	124	10.8	8	12.9

	Willing	592	38.2	142	42.0	430	37.4	20	32.3
	Very willing	743	47.9	171	50.6	540	46.9	32	51.6
	Total	1,551	100	338	100	1,151	100	62	100
Reproductive health	Not at all willing	31	2.0	3	0.9	27	2.4	1	1.7
	Not very willing	49	3.2	6	1.8	41	3.6	2	3.3
	Not sure	143	9.3	27	8.0	107	9.3	9	15.0
	Willing	587	38.0	140	41.4	431	37.5	16	26.7
	Very willing	736	47.6	162	47.9	542	47.2	32	53.3
	Total	1,546	100	338	100	1,148	100	60	100
Medications being taken	Not at all willing	27	1.7	1	0.3	25	2.2	1	1.6
	Not very willing	45	2.9	5	1.5	39	3.4	1	1.6
	Not sure	119	7.7	18	5.3	97	8.4	4	6.5
	Willing	590	38.0	135	39.8	436	37.9	19	30.7
	Very willing	770	49.7	180	53.1	553	48.1	37	59.7
	Total	1,551	100	339	100	1,150	100	62	100
Smoking behaviour	Not at all willing	24	1.6	2	0.6	21	1.8	1	1.6
	Not very willing	28	1.8	3	0.9	24	2.1	1	1.6
	Not sure	67	4.3	12	3.6	53	4.6	2	3.2
	Willing	580	37.5	131	38.9	430	37.5	19	30.7
	Very willing	846	54.8	189	56.1	618	53.9	39	62.9
	Total	1,545	100	337	100	1,146	100	62	100
Alcohol use	Not at all willing	23	1.5	1	0.3	21	1.8	1	1.6
	Not very willing	30	1.9	4	1.2	26	2.3	0	0
	Not sure	67	4.3	16	4.7	49	4.3	2	3.2
	Willing	599	38.8	131	38.8	446	38.9	22	35.5
	Very willing	827	53.5	186	55.0	604	52.7	37	59.7
	Total	1,546	100	338	100	1,146	100	62	100
Illegal drug use	Not at all willing	50	3.2	5	1.5	43	3.8	2	3.3
	Not very willing	47	3.1	6	1.8	40	3.5	1	1.6
	Not sure	141	9.1	31	9.3	109	9.5	1	1.6
	Willing	522	33.9	117	34.9	386	33.7	19	31.2
	Very willing	782	50.7	176	52.5	568	49.6	38	62.3
	Total	1,542	100	335	100	1,146	100	61	100

Table 31: Responses to Question 11 Imagine that the researcher from the study you took part in wants to share your data with other researchers. How willing would you be for them to share anonymised details of your:

5.19 Research Question 2: Does knowing about sharing affect taking part:

5.19.1 Question 15: if you knew your data might be shared, what affect would it have on you taking part in a study?

Respondents were asked whether, if they knew that their data would be shared it would affect their decision to take part in a study (Q15). Of the seven possible answers to this question, there was a slight majority (47.2%) for this knowledge having no effect on their taking part. The next most popular answer was that 'I'd be a bit more cautious about taking part' (38.7%). Very few respondents stated that they would be much more or less likely to take part after learning about sharing (between 3 and 4%) and even fewer (0.8%) reported that they would not take part at all.

Response:	Number of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
I would not take part at all	12	0.8	1	0.3	9	0.8	2	3.3
I'd be much less likely to take part	54	3.6	19	5.7	32	2.9	3	4.9
I'd be a bit more cautious about taking part	586	38.7	111	33.3	456	40.8	19	31.2
It would have no effect on my decision	714	47.2	172	51.7	516	46.1	26	42.6
I'd be a bit more likely to take part	35	2.3	12	3.6	21	1.9	2	3.3
I'd be much more likely to take part	47	3.1	12	3.6	28	2.5	7	11.5
Not sure	65	4.3	6	1.8	57	5.1	2	3.3
Total	1,513	100	333	100	1,119	100	61	100

Table 32: Responses to Question 15 Does knowing about data sharing affect the likelihood of respondents taking part in research?

In secondary analyses, a Chi Square test of independence revealed that source study (χ^2 (d.f.=2, n=1,513) =7.8, p<0.020), age (χ^2(d.f=3, n=1,491) =11.9, p < 0.007), respondents' self-rated health (χ^2 (d.f.=5, n=1494) =16.8, p0.005) and experience of taking part in research (χ^2 (d.f.=6, n=1,366) =46.9, p<0.001) were significantly associated with 'likelihood' of taking part in research after finding out that their data would be shared.

Post-hoc comparisons of age and 'likelihood' and experience of taking part and 'likelihood' identified significant differences in the proportion of respondents who were 'likely' to take part between respondents aged 25-44 and respondents aged 65-74 and respondents who had a 'very positive' experience of taking part compared to those with a neutral experience (neither a positive nor a negative experience). Further details are given in Table 33 below. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
If you knew your data might be shared, what effect would it have on you taking part?	Age 25-44 vs 65-74	0.017
	Very positive (ETP) vs neither	<0.001

Table 33: Q15 Bonferroni correction comparisons- age and experience of taking part.

Respondents from ACONF (58.9%) and the participant groups (57.4%) were almost equally as 'likely' to take part after being made aware of sharing and ALSPAC were only slightly more cautious with (50.5%) reporting that they would continue to take part after being made aware that their data could be shared.

As displayed in Table 34 below, older respondents were more 'likely' to share their data after being informed about sharing with 80% of respondents aged 75-84 being likely to share as compared to 50.5% of those aged 25-44.

	Number of respondents (%)					
Age	**Unlikely**		**Likely**		**Total**	
25-44	551	49.5	563	50.5	1,114	100
45-64	38	45.2	46	54.7	84	100
65-74	112	39.6	171	60.4	283	100
75-84	2	20	8	80	10	100
Total	703	47.2	788	52.9	1,491	100

Table 34: Results for Question 15 by age

Figure 16 below, shows the likelihood of respondents taking part in research after being informed that their data could be shared by self-rated health. Respondents who rated their health as 'excellent', 'good' or 'poor' were more likely than not to agree to take part. Respondents who rated their health as 'very poor' or 'not sure' were markedly less likely than respondents with other health ratings to take part in research after being informed about sharing with only 23.1 and 16.7% respectively saying that they would.

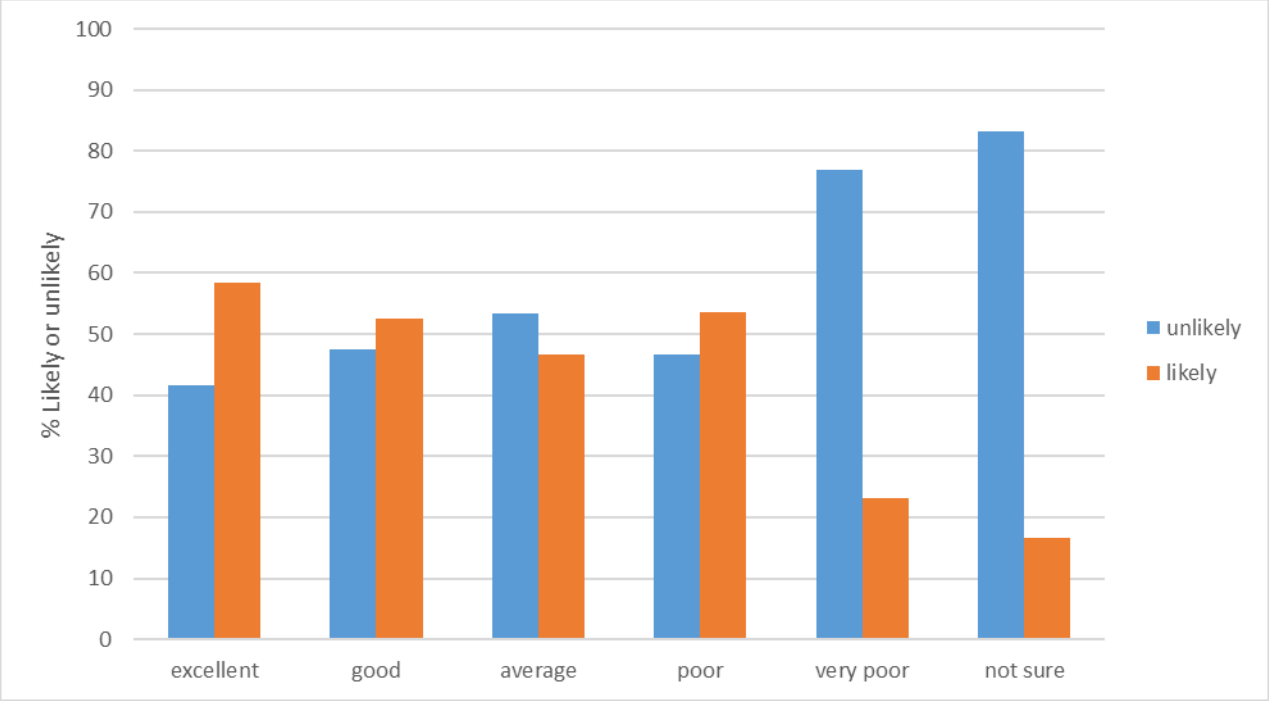


Figure 16: Results for Question 15 by health rating

When observing the percentage of respondents who were still 'likely' to take part after being informed about sharing relative to their experience of taking part in research (Figure 17) it is clear that respondents with a 'negative' or 'very negative' experience of research would be less likely to take part. Respondents who had not taken part or were 'not sure' how their experience was, were approximately equally as 'likely' as 'not likely' to take part in research where their data might be shared. It is fairly clear from Figure 17 that a (very)

positive or neutral (neither positive or negative) experience of research meant that respondents were more likely to take part in research where their data might be shared than those for whom the experience was (very) negative.

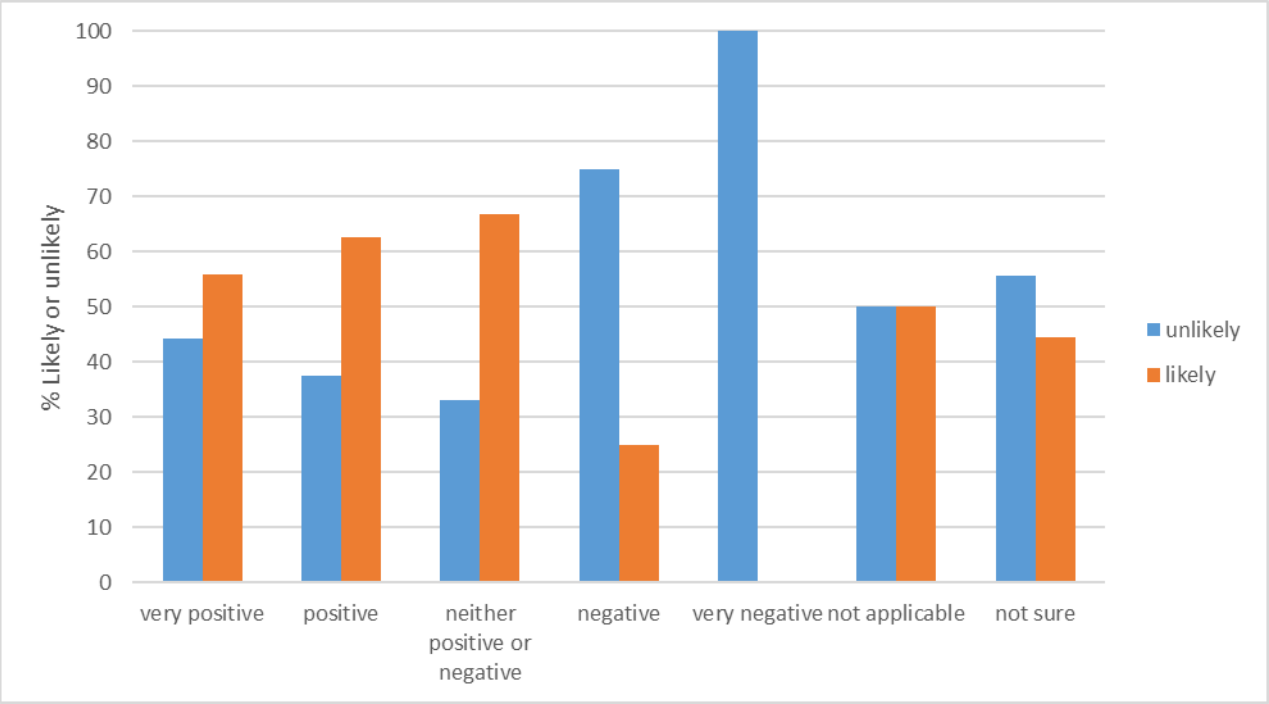


Figure 17: Results for Question 15 by experience of taking part

5.20 Research Question 3: Preferences for sharing:

Questions, 12-14 and 16 to 22 asked respondents for their preferences regarding data sharing processes and procedures such as consent and storage. The results are set out below under the heading of each question.

5.20.1 Question 12: How and when would you like to be asked to share your data?

When asked about their consent *type* preferences, respondents were almost evenly split between agreeing that a single consent at the beginning of the original study could cover all future sharing (39%) and wanting to re-consent each time the data is shared with the option to say no (41.7%). Just 5.4% stated that they were happy for their data to be shared without being consulted at all.

Response:	Number of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Once, on the consent form for the original study	589	39.0	125	37.5	434	38.9	30	49.2
Every time it is shared, with the option for me to say no	631	41.7	119	35.7	490	43.9	22	36.1
Just let me know every time it is shared	135	8.9	29	8.7	104	9.3	2	3.3
There is no need to ask me, just share it	82	5.4	34	10.2	44	3.9	4	6.6
I have no preference	73	4.8	26	7.8	44	3.9	3	4.9
Total	1,510	100	333	100	1,116	100	61	100

Table 35: Responses to Question 12 How and when would you like to be asked to share your data?

In secondary analyses, Chi squared tests of independence revealed that respondents' age, gender, education, source study, experience of taking part and self-rated health were significantly associated with respondents' consent preferences (yes or no to each option presented).

The number of significant associations between respondents' preferences for consent to share varied across these independent variables. Age, gender, source study and experience of taking part in research were the independent variables with the highest number of significant associations with the potential types of consent. There was no significant association between questions 12a and 12c and age. between question 12a and 12e and gender or questions 12c and 12d and experience of taking part. All significant associations for Question 12 can be viewed in Appendix K.

Post-hoc comparisons of age by consent preferences identified significant differences in proportions of respondents answering 'yes' between respondents who were aged 25-44 and 65-74 and respondents aged 25-44 and 45-64. Significant differences in proportions of respondents answering 'yes' were also observed between respondents who were male and female and between respondents who had a (very) positive and negative experience of research. Further details are given in Table 36 below.

Question	Categories	p-value
Every time data is shared with the option to say no	Age 25-44 vs 65-74	0.011
	Male vs female	0.008
There is no need to ask me, just share	Age 25-44 vs 65-74	0.003
	Male vs female	0.011
I have no preference	Age 25-44 vs 65-74	<0.001
	age 25-44 vs 45-64	0.009
	Very positive vs negative	<0.001
	Positive vs negative	<0.001

Table 36: Q12 Bonferroni correction comparisons- age and gender.

All results from post-hoc analyses can be found in Appendix L.

Preferences for consent varied across each independent variable. When it came to source study (see Table 35), ALSPAC respondents were the most likely to prefer to consent each time their data was requested for sharing (43.9%) whilst respondents from ACONF and participant groups expressed a slight preference for a single consent (39.2% and 47.5% respectively).

Results for consent preferences by age, gender and experience of taking part as presented below in Figures 18-20.

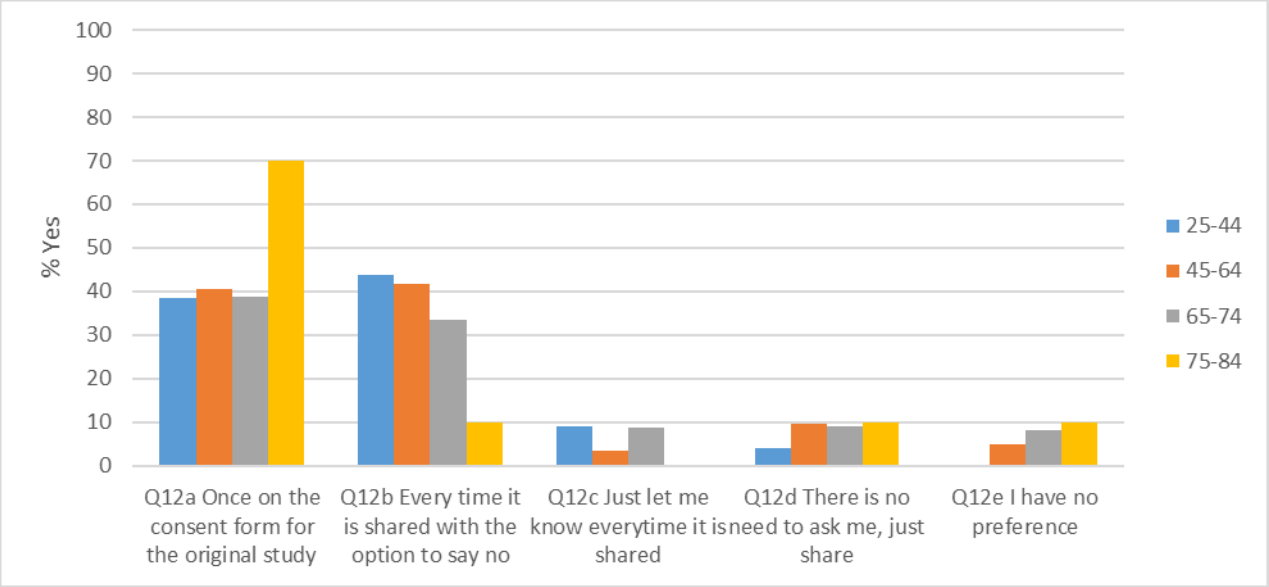


Figure 18: Responses to Question 12 by age

Figure 18 demonstrates that older respondents, i.e. aged 75-84, were vastly more likely, and more likely than respondents of other ages, to choose a single consent for sharing as part of consent for the original study. The youngest respondents, aged 25-44 were slightly more

likely to favour an individual consent model, consenting every time their data was requested (43.8%) as compared to a single consent (38.6%). This fits with the results for source study as most of the age group 25-44 were respondents from ALSPAC and most respondents aged 75-84 were from ACONF. Few respondents from any age group agreed that there was no need to ask or that data could be shared with researchers just letting participants know afterwards.

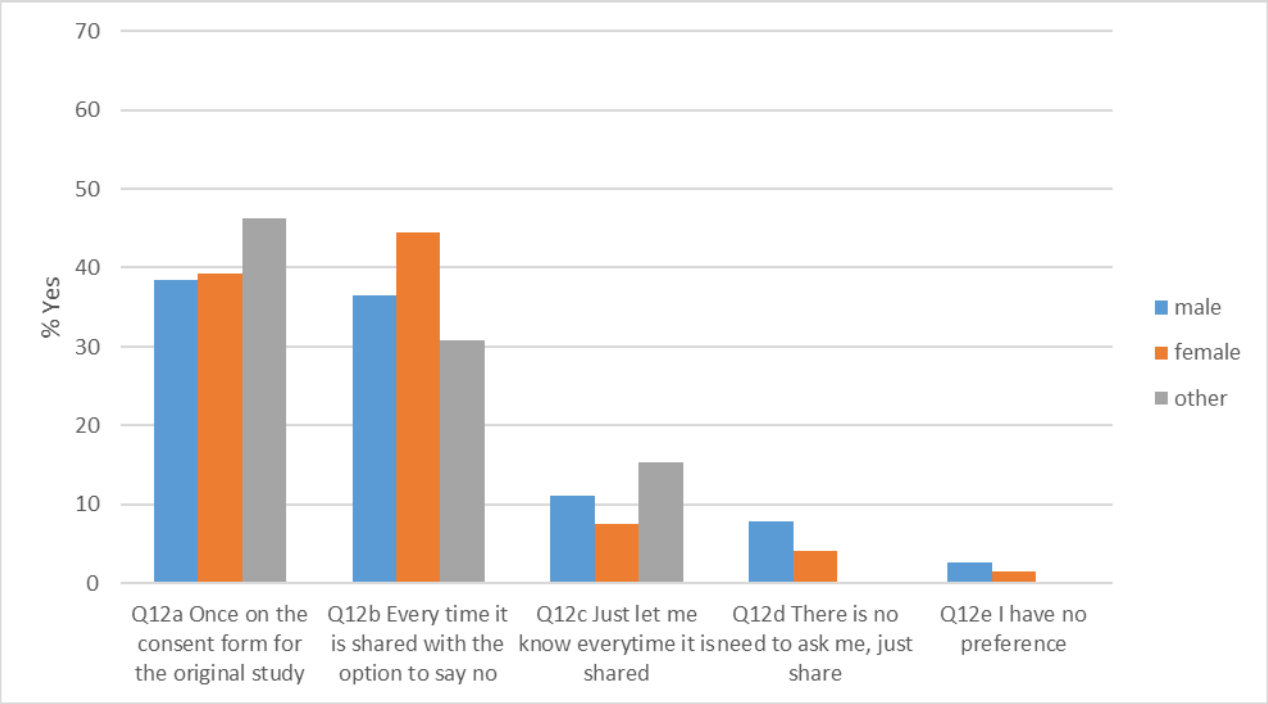


Figure 19: Responses to Question 12 by gender

Female respondents were slightly more cautious than male respondents; they expressed a preference for consenting each time their data was shared, with the option to say no, whilst males expressed a very slight preference for a one-off consent at the time of the original study. Respondents who gave their gender as 'other' were those most likely to agree to a one-off consent (46.2%).

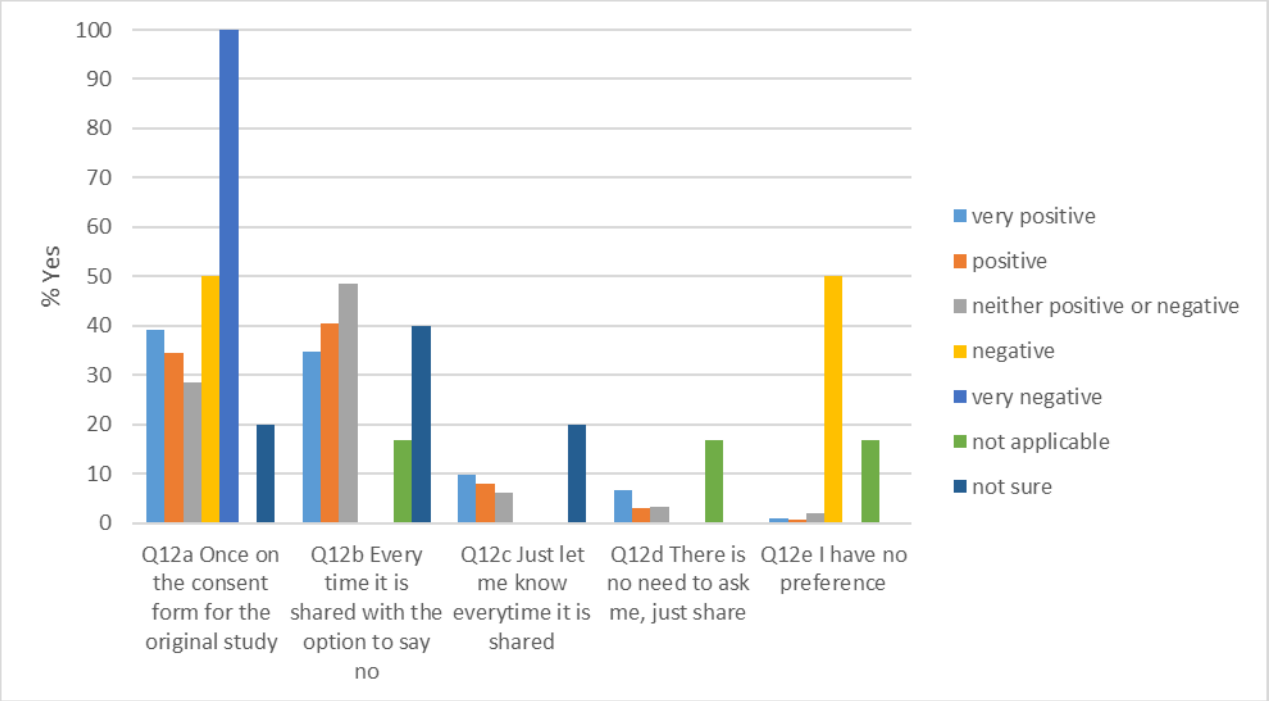


Figure 20: Responses to Question 12 by experience of taking part

Preferences for consent varied by experience of taking part in research with no clear pattern observed. Respondents with a 'positive' experience appeared to prefer to consent each time their data was requested (40.4%) whilst respondents who had a 'very positive' experience were more likely to answer that a one-off consent was adequate (39.12%). However, it is debatable whether there is a great difference between respondents who had a positive versus a very positive experience. There was only one respondent who had a 'very negative' experience of taking part so despite appearances in the bar chart in Figure 20 it is not significant that 100% of those with a 'very negative' experience selected a one-off consent.

5.20.2 Question 13: What information would you like to see on the consent form before you agree to share your data?

Question 13 asked respondents what information (about sharing) they would like to see on the consent form. A high proportion (97.4%) of respondents agreed that the consent form should explain that their data should be shared, and approximately ninety percent of respondents thought that how the researchers would protect respondent's identities, who might benefit from using their data and with whom the data might be shared should also be explained on the consent form. Respondents were less interested in how or where the data would be stored. Only sixteen respondents (1%) thought that none of these things would convince them to share their data.

	Response:	Number of respondents (%)							
		All respondents		ACONF		ALSPAC		PPI groups	
Explain that my data may be shared	Yes	1,471	97.4	321	96.4	1,089	97.5	61	100
	No	24	1.6	4	1.2	20	1.8	0	0
	Not sure	16	1.6	8	2.4	8	0.7	0	0
	Total	1,511	100	333	100	1,117	100	61	100
HOW the researchers will protect (anonymise) my identity.	Yes	1,416	94.0	295	89.4	1,064	95.3	57	93.4
	No	65	4.3	25	7.6	36	3.2	4	6.6
	Not sure	26	1.7	10	3.0	16	1.4	0	0
	Total	1,507	100	330	100	1,116	100	61	100
Explanation of WHO might benefit from using my data	Yes	1,361	90.5	291	88.7	1,017	91.1	53	83.3
	No	91	6.1	25	7.6	62	5.6	4	6.7
	Not sure	52	3.5	12	3.7	37	3.3	3	5.0
	Total	1,504	100	328	100	1,116	100	60	100
Details of WHERE the data will be stored.	Yes	1,138	75.8	238	73.2	852	76.3	48	80.0
	No	254	16.9	56	17.2	188	16.9	10	16.7
	Not sure	109	7.3	31	9.5	76	6.8	2	3.3
		1,501	100	325	100	1,116	100	60	100
Details of HOW the data will be stored.	Yes	1,122	75.0	240	74.3	837	75.1	45	76.3
	No	261	17.5	57	17.7	193	17.3	11	18.6
	Not sure	113	7.6	26	8.1	84	7.5	3	5.1
	Total	1,496	100	323	100	1,114	100	59	100
Details of WHO the data might be shared with.	Yes	1,423	94.5	300	90.6	1,067	95.7	56	93.3
	No	52	3.5	16	4.8	33	3.0	3	5.0
	Not sure	31	2.1	15	4.5	15	1.4	1	1.7
	Total	1,506	100	331	100	1,115	100	60	100

Table 37: Responses to Question 13 What information would you like to see on the consent form before you agree to share your data?

	Number of Respondents (%)							
	All participants		ACONF		ALSPAC		PPI groups	
Q13a None of the above would convince me to share	16	1.0	3	0.8	12	1.0	1	1.6

Table 38: Responses to Question 13a None of the above would convince me to share

In secondary analyses, a Chi Square test of independence revealed that all independent variables (age, gender, education, deprivation quintile, source study, education, experience of taking part and their rating of their overall health) with the exception of ethnicity were significantly associated with information given on the consent form that respondents would like to see before agreeing to share data (yes or no to each option presented).

The number of significant associations between respondents' preferences for consent to share varied across these independent variables. Gender and experience of taking part in research were the independent variables with the highest number of significant associations with consent information. All significant associations for Question 13 can be viewed in Appendix K.

Post-hoc comparisons between gender and consent information identified significant differences in proportions of respondents wanting to see various information types between males and females. Comparisons between experience of taking part and consent information identified significant differences in proportions between respondents who had a 'positive' or 'very positive' experience and respondents who reported a 'negative' experience. Results of comparisons by question are displayed in Table 39. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
Explain that my data may be shared	Neither vs negative	0.041
	Neither vs not applicable	0.041
HOW the researchers will protect (anonymise) my identity	Male vs female	0.003
	Very positive vs negative	0.001
	Positive vs negative	0.001
Explanation of WHO might benefit from using my data	Male vs female	0.009
Details of WHERE the data will be stored	Male vs female	0.032
Details of WHO the data might be shared with	Male vs female	<0.001
	Very positive vs negative	0.001

Table 39: Q13 Bonferroni correction comparisons- gender and experience of taking part.

Proportionally, preferences for consent varied across each independent variable. There were few differences between males, females and those that gave gender as 'other' in selecting information that they would like to see on the consent form. Females were between 1 and 6% more likely than males to select all types of information. Respondents who gave gender as 'other' were very slightly more likely than males (4-9%) or females (2-5%) to require an explanation that data may be shared and how identity will be protected.

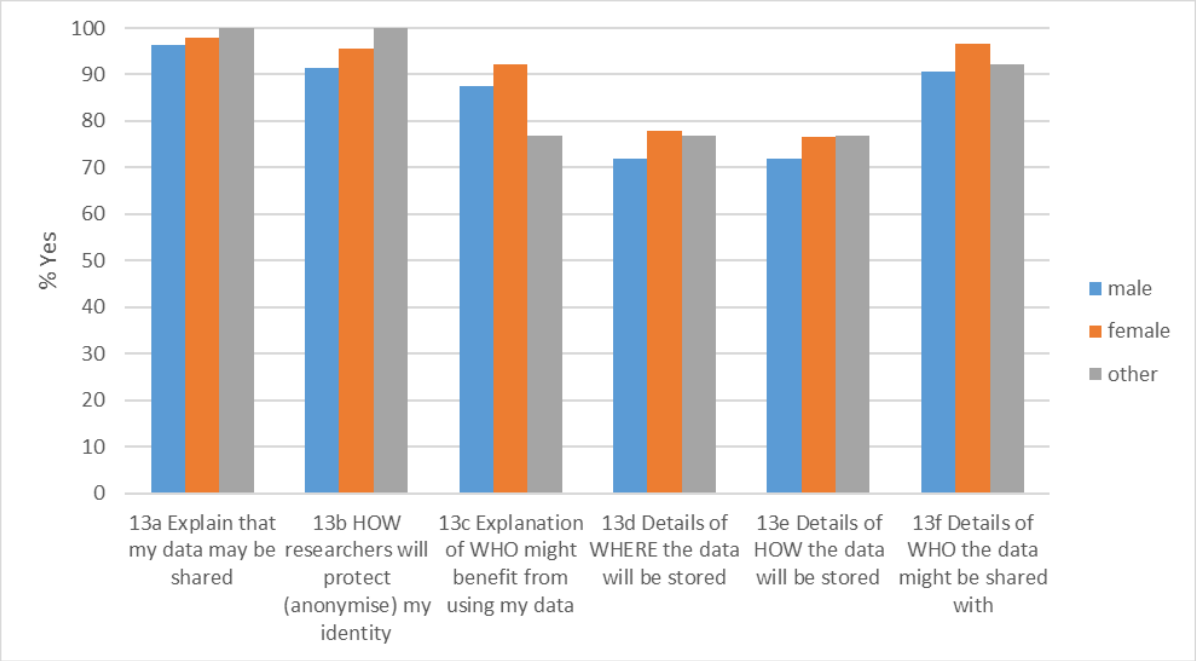


Figure 21: Responses to Question 13 by gender

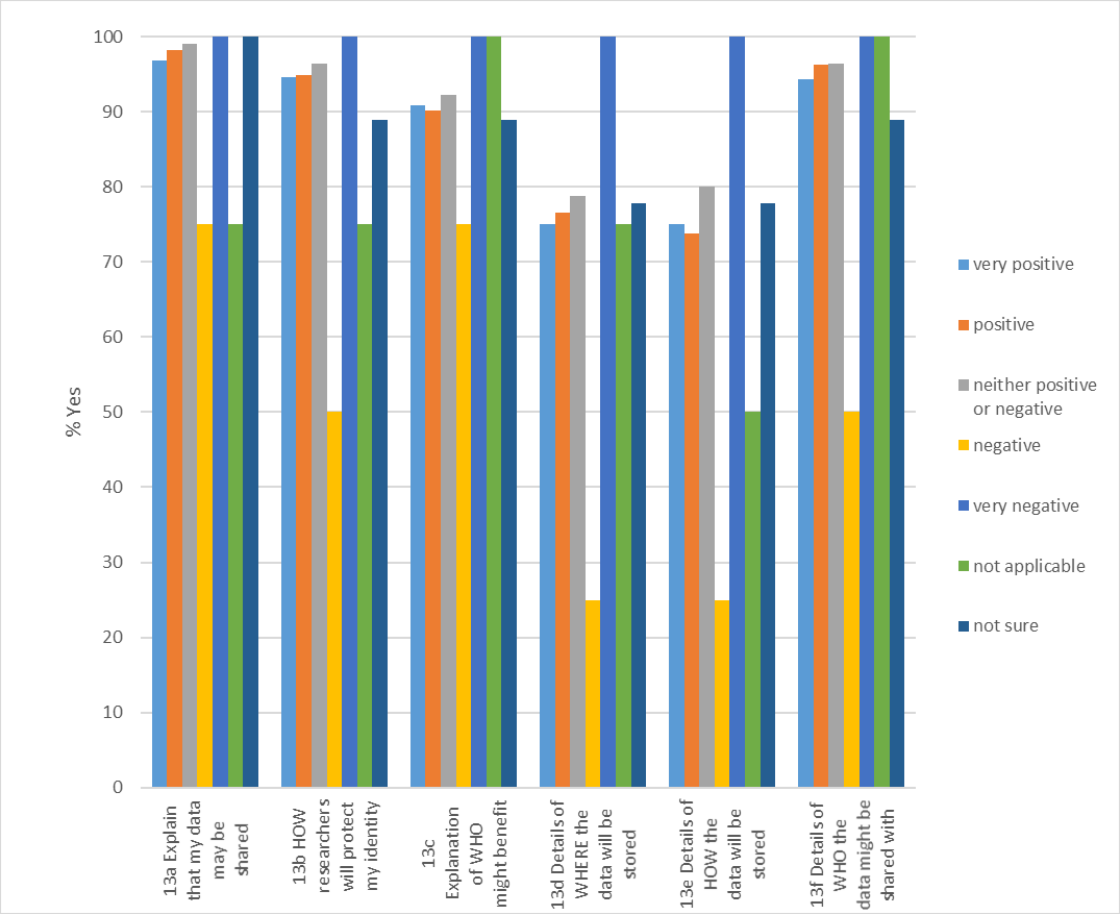


Figure 22: Responses to Question 13 by experience of taking part

Figure 22 shows that most respondents, regardless of their experience of taking part in research thought that all types of information should be included on the consent form, with

slightly less interest in where and how the data will be stored. Again, we must remember that only one individual had a 'very negative' experience so 100% of respondents with a very negative experience selecting each type of information is not significant. In addition, it can be observed that respondents with a 'negative' experience of taking part seem less likely than respondents with other experiences to select all types of information, particularly 'where' and 'how' data will be stored.

5.20.3 Question 14: How important is it that you are informed on the consent form that your study data might be shared?

When asked *how important* it was that they were informed in the consent form that their data might be shared, the majority of respondents (n= 1,002, 66.3%) thought it 'very important'.

Response:	Number of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Very important	1,002	66.3	216	64.9	738	66.0	48	78.7
Somewhat important	399	26.4	84	25.2	308	27.6	7	11.5
Not very important	79	5.2	24	7.2	51	4.6	4	6.6
Not at all important	24	1.6	7	2.1	15	1.3	2	3.3
Not sure	8	0.5	2	0.6	6	0.5	0	0
Total	1,512	100	333	100	1,118	100	61	100

Table 40: Responses to Question 14 How important is it that you are informed on the consent form that your study data might be shared?

5.20.4 Question 16: Would you prefer to give consent separately for each type of organisation your data could be shared with?

When given the chance to state whether they would prefer to give consent separately for each type of organisation that data could be shared with, most respondents (n=907, 60.1%) answered 'yes'.

Response:	No. of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Yes	907	60.1	182	54.8	688	61.5	37	61.7
No	403	26.7	115	34.6	271	24.2	17	28.3
Not sure	200	13.3	35	10.5	159	14.2	6	10.0
Total	1,510	100	332	100	1,118	100	60	100

Table 41: Responses to Question 16 Would you prefer to give consent separately for each type of organisation your data could be shared with?

5.20.5 Question 17: Do you think a register of participants willing to share their study data is a good idea?

Respondents were also asked about whether they thought a register of participants who were willing to share their data was a good idea. Most (n=961, 63.8%) thought that this was a good idea.

Response:	No. of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Yes	961	63.8	218	66.7	706	63.1	37	60.7
No	195	12.9	33	10.1	151	13.5	11	18.0
Not sure	351	23.3	76	23.2	262	23.4	13	21.3
Total	1,507	100	327	100	1,119	100	61	100

Table 42: Responses to Question 17 Do you think a register of participants willing to share their study data is a good idea?

5.20.6 Question 18: Would you be willing to be named on it?

Despite a majority of respondents thinking that a register was a good idea, when asked whether they would be willing to be named on it, only 50.7% (n= 763) agreed that they would.

Response:	No. of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Yes	763	50.7	165	50.5	565	50.5	33	54.1
No	244	16.2	67	20.5	160	14.3	17	27.9
Not sure	499	33.1	95	29.1	393	35.2	11	18.0
Total	1,506	100	327	100	1,118	100	61	100

Table 43: Responses to Question 18 If a register of participants who are willing to share their study data existed, would you be willing to be named on it?

5.20.7 Question 19: How would you prefer your study data to be stored?

Respondents were given a brief statement about storage of data with controlled or open access and then asked how they would prefer their study data to be stored. Unsurprisingly the majority preferred controlled access (n=1.301, 86.9%) with only 3.7% (n=56) preferring open access.

Response:	No. of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Open access	56	3.7	16	4.9	35	3.2	5	8.2
Controlled access	1,301	86.9	289	88.4	958	86.3	54	88.5
No preference	76	5.1	17	5.2	58	5.2	1	1.6
Not sure	65	4.3	5	1.5	59	5.3	1	1.6
Total	1,498	100	327	100	1,110	100	61	100

Table 44: Responses to Question 19 How would you prefer your study data to be stored?

In secondary analyses there were no significant associations between question 19 and the independent variables.

5.20.8 Question 20: If data has controlled access: Who do you think should give permission for data to be shared and used again?

Respondents were then asked who they thought should give permission to share data in a controlled access model. The majority of respondents (n=580, 38.8%) thought that the participants who took part in the study should decide, followed by the organisation where the data was collected (n=349, 23.3%).

Response:	No. of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
The participants who took part should decide	580	38.8	120	36.7	438	39.5	22	37.3
The researcher(s) who collected it	285	19.1	61	18.7	219	19.8	5	8.5
The organisation where the original researcher(s) work	349	23.3	95	29.1	236	21.3	18	30.5
An independent committee	138	9.2	30	9.2	99	8.9	9	15.3
Other	7	0.5	0	0	6	0.5	1	1.7
Not sure	136	9.1	21	6.4	111	10.0	4	6.8
Total	1,495	100	327	100	1,109	100	59	100

Table 45: Responses to Question 20 If data has controlled access: Who do you think should give permission for data to be shared and used again?

In secondary analyses Chi Square tests of independence revealed that age, gender, education and source study were significantly associated with who respondents thought should make decisions about sharing.

The number of significant associations between respondents' preferences for consent to share varied across these independent variables. Gender and source study were the independent variables with the highest number of significant associations with the potential organisations that could make decisions about sharing. All significant associations for Question 20 can be viewed in Appendix K.

Post-hoc comparisons between gender and respondents' preferences for sharing decisions identified significant differences in proportions saying 'yes' between males and females, males and 'other' and females and 'other'. Further details are displayed in Table 46. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
The researcher(s) who collected it	Male vs female	0.008
The organisation where the original researcher(s) work	Male vs female	<0.001
An independent committee	Male vs other	0.039
	Female vs other	0.016

Table 46: Q20 Bonferroni correction comparisons- gender.

As exhibited in Table 45, respondents from each source were almost equally likely to select 'the participants who took part' or 'the researchers who collected it' as the group who should make decisions about sharing. However, we can see that ALSPAC (21.3%) were slightly less likely that ACONF (29.1%) or the participant groups (30.5%) to think that the organisation where the original research took part should make sharing decisions. The respondents from the participant groups were 5% more likely than other respondents to think that an independent committee should make the decision to share.

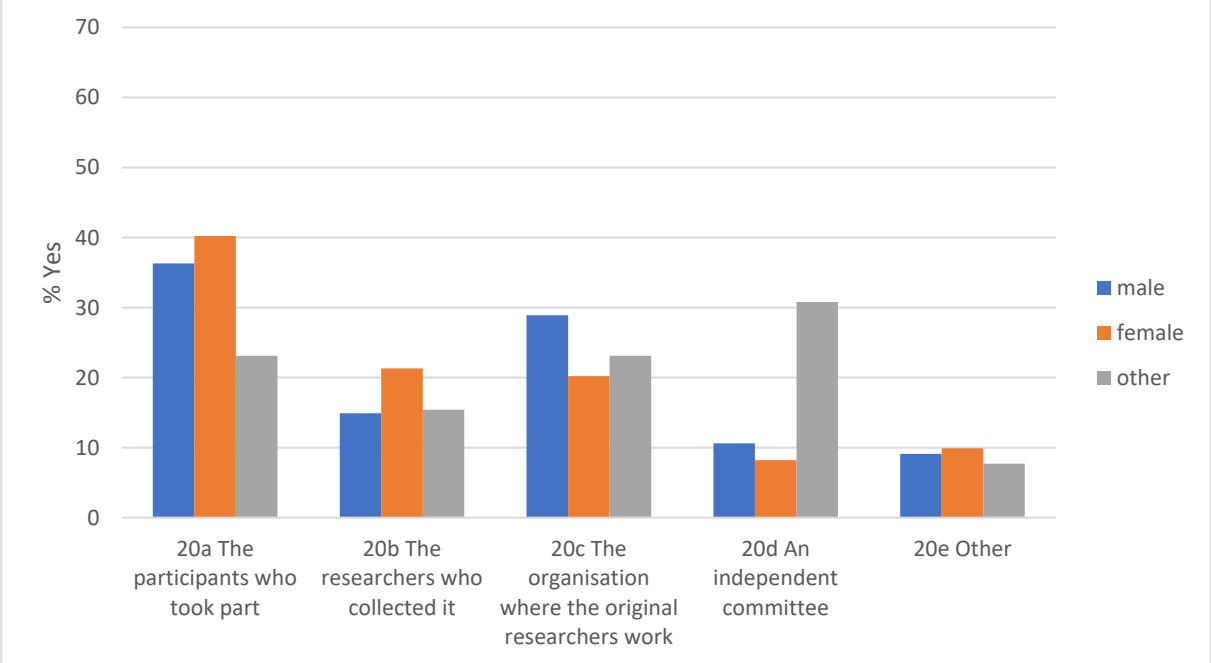


Figure 23: Results for question 20 by gender.

Both male (36.3%) and female (40.2%) respondents were most likely to answer that the participants who took part should make sharing decisions over any other potential organisation. For male respondents the second most popular response was the organisation where the original researchers work (28.9%) but for females, the second most popular option was the original researchers (21.3%). For respondents who gave their gender as other, an independent committee was the most popular choice (30.8%) for sharing

decisions, and these respondents were about 3 times more likely to choose a committee than male or female respondents.

5.20.9 Question 21: Who do you think should 'own' the data collected during a study?

Respondents were asked about who they thought should 'own' the data collected during a study. The majority (n=831, 49.4%) selected 'me/the participants who took part' followed by 'the researcher(s) who collected it' (n=761, 45.2%) and 'the organisation where the original researcher(s) work' (n=742, 44.1%). Only 1.5% (n=26) thought that 'anyone who uses it' should own the data.

Response:	No. of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Me/the participants who took part	831	49.4	208	52.7	616	50.2	7	11.1
The researcher(s) who collected it	761	45.2	133	33.7	624	50.9	4	6.4
The organisation where the original researcher(s) work	742	44.1	194	49.1	543	44.3	5	7.9
Anyone who uses it	26	1.5	6	1.5	20	1.6	0	0
Whoever stores it	30	1.8	6	1.5	24	2.0	0	0
No one	41	2.4	10	2.5	30	2.5	1	1.6
Other	5	0.3	1	0.3	4	0.3	0	0
Not sure	76	4.5	10	2.5	66	5.4	0	0

Table 47: Responses to Question 21 Who do you think should 'own' the data collected during a study?

In secondary analyses Chi square tests of independence revealed that all independent variables analysed (age, gender, ethnicity, education, deprivation and source study) except for self-rated health were significantly associated with at least one of the potential responses for who respondents thought should own data collected during a study.

The number of significant associations between who respondents' thought should own data varied across these independent variables. Age and source study were the independent variables with the highest number of significant associations. All significant associations for Question 21 can be viewed in Appendix K.

Post-hoc comparisons between age and who respondents' thought should own data and , source study and who respondents thought should own data identified significant differences in proportions between most age groups and sources of respondents variously, for each of the questions that had a significant relationship with age. Further details are displayed in Table 48. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
Me/the participants who took part	Age 25-44 vs 75-84	0.026
	Age 45-64 vs 65-74	0.017
	Age 65-74 vs 75-84	0.010
	ACONF vs participant groups	<0.001
	ALSPAC vs participant groups	<0.001
The researcher(s) who collected it	Age 25-44 vs 45-64	<0.001
	Age 25-44 vs 65-74	<0.001
	Age 25-44 vs 75-84	0.002
	ACONF vs participant groups	<0.001
	ALSPAC vs participant groups	<0.001
	ALSPAC vs ACONF	<0.001
The organisation where the original researcher(s) work	Age 25-44 vs 75-84	0.013
	Age 45-64 vs 75-84	0.033
	Age 65-74 vs 75-84	0.027
	ACONF vs participant groups	<0.001
	ALSPAC vs participant groups	<0.001

Table 48: Q21 Bonferroni correction comparisons-age and source study.

As displayed in Table 47, respondents from ALSPAC were most likely to answer that the participants who took part or the researchers who collected the data should own it. For ACONF respondents and those from participant groups it was a bit clearer that the participants that took part should own the data. Responses to Question 21 by age are displayed in Figure 24 below.

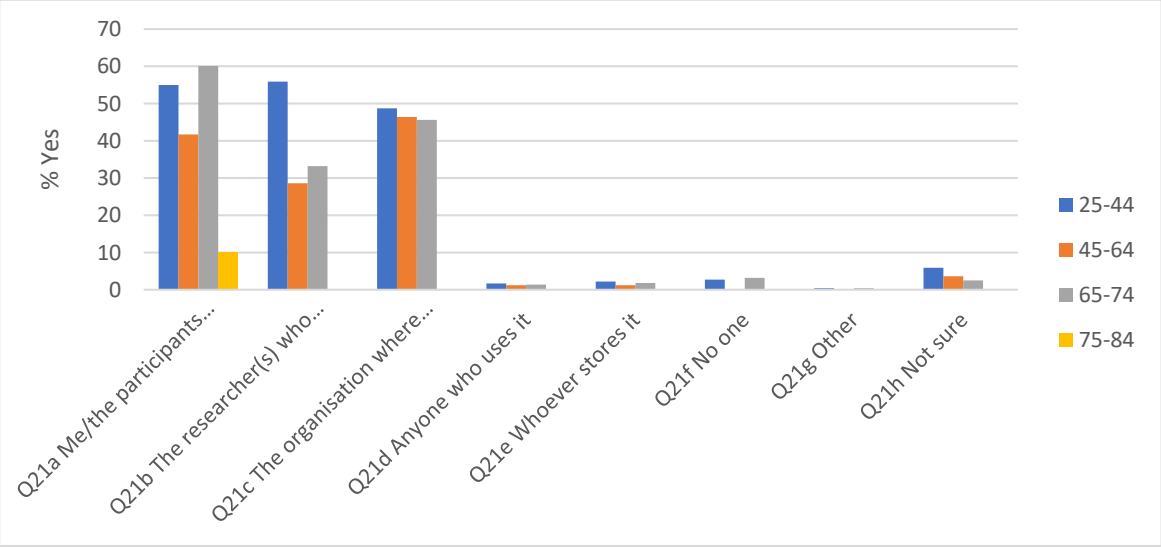


Figure 24: Responses to question 21 by age.

As displayed in Figure 24, respondents aged 25-44 were most likely to answer that the participants who took part or the researchers who collected the data should own collected data. For respondents aged 45-64 there was a slight preference for the organisation who collected the data. Respondents aged 65-74 were more clear that their preference was for the participants who took part to own the data. The oldest respondents preferred that the participants who took part were the ultimate data owner.

5.20.10 Question 22: Do you think it is important that researchers using shared data give feedback telling participants how their data was used?

Finally, the questionnaire asked respondents whether they thought that feedback on how their data was used in secondary research was important. A large majority of respondents (n=1,226, 81.9%) thought that feedback was 'important'. The remaining respondents thought that feedback was either not important (n=126, 8.4%) or that they were not sure (n=145, 9.7%).

Response:	**No. of respondents (%)**							
	All respondents		**ACONF**		**ALSPAC**		**PPI groups**	
Yes	1,226	81.9	244	74.9	929	83.7	53	86.9
No	126	8.4	48	14.7	74	6.7	4	6.6
Not sure	145	9.7	34	10.4	107	9.6	4	6.6
Total	1,497	100	326	100	1,110	100	61	100

Table 49: Responses to Question 22 Do you think it is important that researchers using shared data give feedback telling participants how their study data was used?

In secondary analyses Chi squared tests of independence identified that respondents' age (χ^2(d.f=3, N=1,491)=12.28, Pr = 0.006) and source (χ^2(d.f=2, N=1,497)=14.37, Pr = 0.001) were significantly associated with respondents' view of the 'importance' of feedback (response categories of 'yes' and 'no').

Post-hoc comparisons between age and whether respondents' thought feedback was important identified significant differences in proportions between respondents aged 25-44 and 45-64. Significant differences in proportions of respondents who though feedback was important were also observed between respondents from ACONF and ALSPAC. Further details are displayed in Table 50. All results from post-hoc analyses can be found in Appendix L.

Question	Categories	p-value
Do you think it is important that researchers using shared data give feedback to participants	Age 25-44 vs 45-64	0.027
	ACONF vs ALSPAC	0.001

Table 50: Q22 Bonferroni correction comparisons-age and source study.

As displayed in Table 51, younger respondents (25-44) were between 3 and 12% more likely to think that feedback was important than older age groups (45-64, 65-74 and 75-84) and half as likely as those in the next age group up (45-64) to state that feedback was not important. It is not clear from this result whether the greater desire for feedback is actually due to age, or due to the source study, as all respondents from ALSPAC were in this age group.

	No of respondents (%)					
Age Group:	**No**		**Yes**		**Total**	
25-44	181	16.2	934	83.8	1,115	100
45-64	24	28.6	60	71.4	84	100
65-74	63	22.3	219	77.7	282	100
75-84	2	20	8	80	10	100
Total	270	18.1	1,221	81.9	1,491	100

Table 51: Responses to Question 22 by age group.

5.20.11 Question 29: Do you have any further comments about data sharing or about this survey?

At the end of the survey respondents were given the option to provide a comment either on the experience of taking part in the survey, or on data sharing itself and 196 (11.6%) did so. Responses were split between comments about taking part in the survey, comments directed at ALSPAC or ACONF and comments about data sharing. Those about sharing are further categorised as displayed in Table 52 below. The majority of respondents who decided to comment were remarking about the type of secondary use or recipients that they would prefer in an ideal scenario, which I termed 'conditional consent'. Second to this were comments about being happy for their data to be shared for secondary use with the caveat that they must not be identifiable. A number of respondents commented about the questionnaire itself- whether they found it easy to answer the questions, or how their views or experience influenced the way in which they answered the questions. Some respondents (n=7) addressed comments or questions about sharing, but not necessarily about the survey itself to ALSPAC. Other comments that did not fit into any particular category largely focussed on aspects of data sharing or research not relevant to the questionnaire.

Nature of comments:	No. of respondents (%)							
	All respondents		ACONF		ALSPAC		PPI groups	
Respondent would prefer conditional consent based upon secondary use or recipient	42	21.5	8	17.8	32	24.8	2	9.5
Respondent is happy with sharing as long as they can not be identified	35	17.9	7	15.6	25	19.4	3	14.3
Comments about the questionnaire itself or taking part in the survey	26	13.3	5	11.1	17	13.2	4	19.0
The benefits of sharing/altruistic reasons for sharing	18	9.2	8	17.8	6	4.7	4	19.0
Other comments	16	8.2	7	15.6	8	6.2	1	4.8
Respondent would be interested in feedback on use of their data	13	6.7	2	4.4	7	5.4	2	9.5
Respondent has some reservations about sharing data	13	6.7	2	4.4	10	7.8	1	4.8
Respondent trusts researchers to make appropriate sharing decisions	11	5.6	0	0	11	8.5	0	0
General remarks about consent	8	4.1	1	2.2	3	2.3	4	19.0
Comments or questions addressed to ALSPAC	7	3.6	0	0	7	5.4	0	0
Respondent has no concerns about data sharing	5	2.6	3	6.7	2	1.6	0	0
All comments	195	100	45	100	129	100	21	100

Table 52 Summary of Respondent comments

5.20.12 Significant results summary

Due to the number of variables in the questionnaire, and subsequent volume of significant results from secondary analyses it was difficult to identify the associations between the dependent and independent variables that occurred under the umbrella of research questions one to three (RQ1-RQ3, section 5.7, page 169); respondents' attitudes towards sharing, their preferences for it and how knowing about sharing affects their likelihood of taking part in research. To identify any patterns in the results, the number of significant Chi-squared results attributed to each independent variable included in the secondary analyses were summarised as displayed in Table 53 below:

No. of significant associations	Independent variable								
	Source	Gender	Age	Ethnicity	Education	Health	Deprivation	Experience of taking part	Total no. of significant associations
Total number of significant associations	43	29	36	4	22	26	11	43	214
% of significant associations - total	**20**	**12.1**	**16.7**	**1.4**	**10.2**	**12.1**	**5.1**	**20**	**100%**
No. of significant associations - attitudes to sharing (RQ1)	29	15	25	3	13	20	6	30	141
% of significant associations - attitudes to sharing (RQ1)	**20.6**	**10.6**	**17.7**	**2.1**	**9.2**	**14.2**	**4.3**	**21.3**	**100%**
No. of significant associations - effect on taking part (RQ2)	1	0	1	0	0	1	0	1	4
% of significant associations - effect on taking part (RQ2)	**25**	**0**	**25**	**0**	**0**	**25**	**0**	**25**	**100%**
No. of significant associations - preferences for sharing (RQ3)	13	14	10	1	9	5	5	12	69
% of significant associations - preferences for sharing (RQ3)	**18.8**	**20.2**	**14.5**	**1.4**	**13.0**	**7.2**	**7.2**	**17.4**	**100%**

Table 53 Summary of patterns of independent variables relationships with attitude

As summarised in Table 53 above, there were a total of two-hundred and fourteen significant associations resulting from the secondary analyses, the majority of which (n=43, 20%) were due to the independent variables 'experience of taking part' and 'source study'. The variable with the second largest number of associations was respondent age to which 16.7% of significant associations were attributed.

When looking specifically at research question 1 (RQ1), respondents' attitudes towards sharing, the independent variables with the highest number of associations were again: experience of taking part in research (n=30, 21.3%), source study (n=29, 20.6%) and age (n=25, 17.7%), in that order.

However, when it came to RQ2, there were only four significant associations in total, accounting for 25% of associations with the independent variables each. These variables were source, age, self-rated health and experience of taking part.

For RQ3 variables the independent variables with the most significant associations were gender, source study and experience of taking part. Gender had 14 significant associations (20.2%), source study had 13 (18.8%) and experience of taking part had 12 (17.4%).

Although there were numerous significant associations identified between the independent and dependent variables, the pattern of these significant associations was not always uniform. Using gender as an example, we can see that there were twenty-nine significant associations between gender and research questions one and three. If we examine the secondary analyses for significant associations between gender and the dependent variables in RQ1 (Q5- Q11), we can see that it is not always the same gender (male, female, other) which is significantly associated with either a positive or negative attitude towards sharing. For example, sometimes we identify that those respondents who gave gender as 'other' are more 'concerned' (e.g.: Q6b) and sometimes females are more 'concerned' (e.g.: Q7b). If we ignore only the significant associations and look at all cross-tabulation results for age and dependent variables in RQ1 the lack of a distinct pattern remains. These results can be viewed in Appendix K.

Significant results for the independent variables that had the highest number of significant relationships with dependent variables (e.g., experience of taking part; age) were corrected using the Bonferroni method. The results of these post-hoc tests indicated that there were

significant differences in response patterns between participant groups. For example, respondents who had a 'very positive' or 'positive' experience of taking part in research were often significantly different to those who had a 'negative' experience or stated that they were 'not sure' how their experience was. It is not clear whether respondents answered that they were 'not sure' because they did not wish to answer that question, or genuinely were not sure how they felt about research. This difference between respondents was apparent for comparisons of attitude that fell under the categories of research questions one, two and three. The same was true for gender and age although corrected comparisons for self-rated health only had significant differences in attitudes for research question one- attitudes towards sharing and not, preferences for processes (research question 3).

Generally, though, it appears that a respondents' experience of taking part in research and their age, were most likely to be associated with respondents' attitudes and preferences. Ethnicity and deprivation quintile had the fewest significant relationships with the dependent variables.

5.21 Chapter summary

This chapter has described in detail the process by which the questionnaire survey was distributed, the results collated and analysed as well as the results of these analyses. The next chapter (Chapter 6: Discussion and recommendation for best practice) will first summarise and compare the results of each data collection method used in this study and then go on to examine the strengths and limitations of the study as a whole, identify areas that require further research or clarification and finally, make recommendations for future best practice.

Chapter 6 Discussion and recommendations for best practice

6.1 Introduction

This study aimed to identify research participants' attitudes towards research data sharing via a questionnaire survey, supported by data from a systematic review of existing literature, and to relate the findings from these two strands to current best practice guidance on research data sharing in the UK. This chapter concludes the PhD study, drawing together the three strands of research and providing recommendations for data sharing research and practice, based on the evidence from these sources.

First, I present summarised findings from the systematic review, the grey literature review, and the questionnaire survey before comparing these findings with 'triangulation', exploring whether there were any areas of common ground and discussing the implications of these findings for data sharing research and practice. The comparison of findings is discussed within the context of the original research aims specified in the background chapter (Chapter 1) using the section of the questionnaire survey (and grey literature review) as headings. I then remark upon the successes and limitations of the PhD study, before concluding with my evidence-based recommendations for best practice and identification of topics that require further research or clarification.

6.2 Summary of systematic review findings

The systematic review of existing international literature on the attitudes of research participants and members of the public addressed research questions 1 to 3 and identified 18 relevant papers which were analysed using thematic synthesis. This resulted in the identification of six themes: 1) benefits of data sharing, 2) fears and harms, 3) data sharing processes 4) relationship between participants and research, 5) willingness to share and 6) conditions and pre-requisites for sharing. These themes are summarised briefly below.

Benefits of data sharing: Participants identified three main types of benefit resulting from sharing: benefit to participants or immediate community; benefits to the public more generally; and benefits to science or research. Ideally, they wanted to see benefits to the community, themselves, or future generations if they were to share their data but were still willing to do so without reaping any immediate benefit to themselves.

Fears and harms: When prompted, participants were able to identify potential pitfalls for themselves (and to a lesser extent for science and researchers) related to sharing.

Participants were primarily concerned about being identified in shared data, having their data stolen or receiving unwanted contact from companies. Other consequences such as misinterpretation of data (either deliberately or accidentally) and bias in results were also identified.

Data sharing processes: Participants discussed data sharing processes such as consent and data governance. Consent was viewed as an opportunity to inform participants about sharing. Data stewards or committees with lay members were seen as a way of making participants feel more comfortable with sharing, and of reducing the need for re-consenting. Participants were more likely to share if they understood the process, the benefits were clear, risks were mitigated, and the research was in keeping with participants' values. Some participants also identified more researcher-specific barriers to sharing such as resource implications and the relative novelty of sharing.

Relationship between participants and research: Not all participants were aware of data sharing, and participants thought that sharing processes could be more transparent or better explained to participants by researchers. Although there was uncertainty about what might happen to their data if they were shared, participants generally trusted researchers to make those decisions for them. This trust extended to researchers choosing appropriate projects with which to share the data; for example, participants would prefer that their data was not used for commercial gain. Participants were keen for feedback from researchers about how their data had been used in secondary research and made suggestions for how this might be facilitated.

Willingness to share: Some participants had no concerns regarding sharing, commenting that they expected it to happen, and it was no different to sharing information on social media. Participants discussed data items that they would and would not be comfortable sharing, which could also depend on with whom those data would be shared; universities, scientists and not-for profit organisations were more acceptable than companies who might profit from their data.

Conditions and Pre-Requisites: Participants' consent preferences in respect of data sharing were broadly in line with their overall contact preferences in that, although being re-contacted each time their data was shared would be a welcome courtesy, participants recognised that this could be bothersome for them and impractical for researchers. Sharing could go ahead so long as participants had experienced a full and transparent consent

process during which assurances about anonymisation should be provided. Secure storage and privacy of data was important for most participants, as was knowing who would access it.

6.3 Summary of grey literature review findings

The grey literature review addressed the fourth research question and identified 16 relevant guidance documents from funders or other organisations who facilitate research, published between 2002 and 2019. The guidance from these documents primarily concerned sharing of anonymised research data and was summarised into four main categories. These categories were taken from the 'Data Sharing Processes' and 'Conditions and Pre-requisites' themes of the systematic review. The findings are briefly as follows:

Consent: Eleven of the included documents provided guidance on consent to ensure that it was not only ethical but lawful. Broadly, consent for sharing needs to be considered from the trial set up, and a data sharing statement should be included in the consent form so that consent can truly be informed, and sharing can be ethical. The type of consent given should be considered again at the end of the study when preparing data for sharing. Guidance is also provided on what researchers should do where no consent exists, when research participants wish to withdraw consent and the most appropriate types of consent to apply.

Storage: All but one of the guidance documents referred to storage of data for sharing. Ideally, where data will be stored should be considered at the beginning of a research study to comply with the FAIR principles[8] (Wilkinson et al., 2016) and so that the storage location can be communicated to participants. Guidance on anonymisation is provided by organisations so that participant privacy can be protected but must be balanced with future data utility. Some suggestions are made for secure release of and access to data, for example through use of repositories with restricted access or via data safe havens.

Access to data: All included guidance documents referred to access to data. Guidance was provided on types of access to data (open or controlled), and the ideally transparent processes involved in providing researchers with access to research data, for example reviewing requests for data using independent data access committees to enhance transparency. This section also provided an overview of guidance on preparing data for

[8] To be Findable, Accessible, Interoperable and Reusable.

sharing with sufficient metadata, and formally recording the sharing process with data access agreements.

Type of sharing: Five of the included documents referred to providing feedback to research participants on how their data was used in secondary research to increase transparency. Trust in research and researchers is referred to only briefly, but data should only be shared with bona fide researchers. Sharing with commercial organisations is not precluded by the guidance documents but proposed research should be in the public interest. Appropriate recognition for the original research team is also seen as appropriate and important by eight of the included guidance documents.

6.4 Questionnaire results summary

The questionnaire survey addressed research questions one to three. Responses demonstrated that respondents were open to data sharing and exhibit low levels of concern when asked to share most types of data, but when prompted, respondents were able to express concern or state their preferences for procedures for consent, storage and sharing.

Attitudes towards sharing

Most respondents were 'not at all concerned' about sharing their data with various organisations, and most indicated that they were 'very likely' to agree to share their data for most of the suggested purposes, with respondents more willing to share with universities or hospitals than with pharmaceutical companies or the government. The majority were also able to identify potential benefits of sharing and motivators to share, with a chance to help others, assurance of anonymity, and understanding the use of the data being the most frequently endorsed. A majority of respondents were very willing or willing to share most of the suggested details about themselves but were more likely to be unwilling to share details of mental health or employment.

Does knowing about sharing affect taking part?

Crucially, when asked whether, if they knew that their data would be shared, it would affect their decision to take part in a primary study, the most common answer was that this knowledge would have no effect on their taking part in research.

Respondent preferences

The majority of respondents thought it 'very important' that the consent form lets participants know their data will be shared but were less interested in how or where the data will be stored. When asked about their consent type preferences, respondents exhibited a slight preference for wanting to re-consent each time the data are shared, with the option to say no. Few respondents stated that they were happy for their data to be shared without being consulted at all, and those that did were significantly more likely to be aged 45-85+ and in poor health.

Most respondents thought a register of participants who were willing to share their data was a good idea, but fewer were willing to be named on such a hypothetical register. Statistically speaking, only respondents with self-rated average health were more likely to think that a register was a good idea but again, were less likely to want to be named on such a register. The vast majority of respondents preferred controlled access to data and thought that the participants who took part in the study should decide whether access to data should be granted, closely followed by the organisation where it was collected.

Almost half of respondents thought that the participants themselves should own the data whilst a comfortable majority felt that feedback on how their data was used was 'important'.

Significant results

Through secondary analyses over two-hundred significant associations were identified. Respondents' experiences of taking part in research, source study and their age were the independent variables most likely to be associated with attitudes towards sharing. These results are discussed in more detail below.

6.5 Commentary on results from all sources

This section attempts to draw together the results of the systematic and grey literature reviews and the questionnaire survey and to identify the degree of triangulation between them. Where possible there will also be reference to other literature that supports the findings of this study, but there is little literature regarding attitudes towards sharing of data from health research, longitudinal or clinical trials studies that is not already included in this study, and so the primary focus will be upon the triangulation.

The aim of this study has always been to conduct a questionnaire survey and so the survey results are considered the main data source. Below, data from the questionnaire are referred to as coming from 'questionnaire respondents' or 'respondents', while data from the systematic review are referred to as coming from 'systematic review participants' or 'participants' and the data from the grey literature are referred to as 'grey literature' or 'guidance'. I identify and discuss whether the views expressed by the respondents to the questionnaire are reflected by the participants in the literature review, and whether the guidance documents identified during the grey literature review reflect the views or concerns of respondents and participants.

6.5.1 What are participants attitudes towards data sharing?

The questionnaire survey measured respondents' attitudes towards sharing through questions 5 to 11 and via example sharing scenarios. Prior to the questions, respondents were given some explanatory statements about types of research and data sharing, including anonymisation, consent and access options. Respondents were reminded that data would be anonymised prior to sharing.

Concern about sharing

First, respondents were asked how concerned they would be if they found out that a study in which they were involved was sharing their data. The most common response was 'depends who it is shared with' which is reflective of findings in respect of participants identified in the systematic review, who were happy to share data but with prior stipulations about who could use it (Mello *et al.*, 2018; Shah *et al.*, 2018; Mozersky *et al.*, 2020).

For example, participants were more likely to agree to share with university scientists than with commercial companies (Mello *et al.*, 2018) and were least likely to want to share with drug companies (Shah *et al.*, 2018) whilst participants in Europe expressed a preference for their data remaining with European researchers (Shah *et al.*, 2018).

It is entirely possible that had respondents to my survey not been provided with the response option of 'depends who it is shared with' statement, they would have instead selected 'not very concerned' which was the second most frequently endorsed category. Respondents being 'not very concerned' about sharing tallies with findings from most literature regarding attitudes towards sharing of health data which state that levels of

concern are low, although specific harms are identified when prompted (Clerkin *et al.*, 2013; Courbier *et al.*, 2019; National Academies of Sciences and Medicine, 2020).

The grey literature does not really consider how concerned research participants might be about sharing, but some documents explain how to counter potential concerns, and refer to the trust placed in researchers to oversee secondary research whilst respecting the original providers of the data and maintaining this trust (Lowrance, 2002; Medical Research Council, 2011; Medical Research Council, 2017). Researchers should open a *"dialogue with the public"* whereby trust is synonymous with consent for sharing and by which researchers can identify research participants' preferences for and concerns about sharing and come to a consensus about procedures such as protection of data or feedback on its use (Lowrance, 2002, p. 66).

Concern about who data is shared with

Question six of the questionnaire invited respondents to elaborate on how concerned they would be if they found out that their data were being shared with specific types of recipient such as researchers at the organisation where their data was collected, researchers at universities, pharmaceutical companies, the government or was freely available on the internet. Most respondents answered: 'not at all concerned' to questions about sharing in respect of most of the presented options, but respondents did exhibit concern regarding the potential for sharing data 'on the internet for anyone to use' (by which I was implying a completely open access repository) with 61% (n=956) of respondents 'very concerned' about this. We cannot be absolutely certain that this concern is because respondents want to know who will be using their data, rather than being related to concerns about something else, such as privacy. Although respondents were told that data would be anonymised before any sharing took place, the degree to which they understand anonymisation or recalled this information when answering may have influenced responses about their data being available online. In the context of this question, we can assume that respondents would prefer to know who is using their data and that having it freely available on the internet makes them uneasy.

The systematic review identified that some research participants would be encouraged to share by knowing who was going to be using their data or knowing that they were appropriately qualified (Hate *et al.*, 2015; Manhas *et al.*, 2015; Manhas *et al.*, 2016; Mozersky *et al.*, 2020); it also showed that others would like the recipients of their data to

be recorded on an on-going basis (Shah *et al.*, 2018). Participants' preference for caution in choosing who is receiving their data is only echoed in the grey literature by two documents. Open access is a potential issue for The Academy of Medical Sciences, who cited concerns about privacy and re-identification (The Academy of Medical Sciences, 2013) and Tudur-Smith *et al* who reflected on the risk of re-identification as well as several other researcher-centric issues. The grey literature said little about participants' concern when sharing with different organisations but the point about communicating potential types of uses or who data might be shared with is picked up by three documents (Corti *et al.*, 2014; Tudur-Smith *et al.*, 2015; Castell *et al.*, 2018).

Concern about potential harms

Question 7 of the questionnaire asked respondents which potential harms arising from sharing they would be most concerned about. Primarily, respondents' concerns related to privacy such as being identified or having their data stolen. The systematic review also found the most commonly raised concerns to be re-identification, misuse of data (including theft) as well as use of data for profit (Asai *et al.*, 2002; Cheah *et al.*, 2015; Hate *et al.*, 2015; Manhas *et al.*, 2015; Merson *et al.*, 2015; Mello *et al.*, 2018; Mozersky *et al.*, 2020). Misuse of data could refer to data being misinterpreted (either through misunderstanding or to suit a particular purpose) or being used for a purpose other than that originally agreed such as unwanted contact or a study that did not align with participants' values (Manhas *et al.*, 2015; Merson *et al.*, 2015; Colombo *et al.*, 2019). It was therefore important for some of the participants represented in the systematic review that data was shared only for purposes that they had agreed in advance. The questionnaire respondents were less concerned about their data being used to make a profit instead of being used for research purposes (44% 'very concerned') than they were about being identified (64.7% 'very concerned').

Privacy concerns were addressed in the grey literature, where a lot of information was provided for researchers regarding preserving privacy of data, for example through anonymisation (Lowrance, 2002; Corti *et al.*, 2014; Tudur-Smith *et al.*, 2015; Medical Research Council, 2017) with specific steps that can be taken to anonymise provided by the MRC, Corti *et al* and Tudur-Smith *et al.* It is not clear whether privacy and anonymity is so thoroughly covered in the grey literature because it is participants' primary concern or because it is simply the easiest for researchers to address as one of the fundamentals of managing and sharing data.

Some questionnaire respondents reported that they would be concerned about researcher issues such as lack of recognition for the original research team and the potential for the act of sharing stopping researchers from doing their own original research, but it might be that these harms were selected as concerns simply because the options were provided. I've also since identified that the response 'if it stopped researchers doing their own original research' is somewhat ambiguous, and respondents could have interpreted this differently to my intended meaning, which was that researchers would conduct secondary research instead of their own primary research. Respondents might have thought the requirement to share meant that researchers would be deterred from conducting research. However, both of these are still researcher-centric concerns that some respondents selected as important to them.

The systematic review identified that some participants were able to identify potential harms that were researcher orientated. These included secondary results being biased if data was misinterpreted by researchers who were not familiar with the dataset or originating study (Mozersky *et al.*, 2020), or *"poor quality science"* being conducted by secondary researchers (Mello *et al.*, 2018, p. 2206). These concerns seemed less concerning to participants than the potential harms that they identified in relation to themselves. Unsurprisingly the grey literature provided ample consideration of researcher issues, but of those identified by participants, only researcher recognition (Medical Research Council, 2011; ESRC, 2015; Tudur-Smith *et al.*, 2015; HEFCE *et al.*, 2016; Cancer Research UK, 2017; UK Research and Innovation, 2018; NIHR, 2019) and misinterpretation of data were explicitly identified by the included guidance (Medical Research Council, 2011; Tudur-Smith *et al.*, 2015; Open Research Data Task Force, 2018). In the grey literature researcher recognition was mentioned as a stipulation for sharing rather a potential harm if it was lacking. Due consideration was given to the importance of providing adequate metadata, context and sharing with appropriately qualified secondary researchers to avoid misinterpretation. Although these supplementary documents may seem to have nothing to do with participants, appropriate use of metadata and supporting information such as case report forms, syntax files and protocols could reduce the possibility of misinterpretation of data, which was a concern for systematic review participants (Mozersky *et al.*, 2020) and a number of questionnaire respondents (44% were 'very concerned').

I incorporated issues such as misinterpretation, lack of recognition for original researchers and researchers not doing their own original research into the questionnaire because they had been explored in the systematic review. However, participants in the systematic review were prompted to discuss these issues by researchers themselves, and it makes sense that these are not key or spontaneous concerns for participants or questionnaire respondents. By discussing and emphasising potential researcher concerns, participants demonstrate that they would prefer that their data are used well, maximising the benefits of sharing in the first place, as called for by journals and funders (Mello *et al.*, 2013; Carr and Littler, 2015; Loder and Groves, 2015). There is no point in participants risking their health, giving up their time or allowing their data to be shared if it is not used appropriately.

Likelihood of giving permission for data to be shared

In Question 8 respondents were asked how likely they would be to give permission for their data to be shared for various purposes. The majority of questionnaire respondents indicated that they were 'very likely' to agree to share their data for research in a university, hospital, or to inform the public about a health issue. This is reflective of the systematic review participants, who expressed a preference for sharing with *"qualified researchers"* chosen because of their credentials (Manhas *et al.*, 2015). As with the questionnaire respondents, those in the systematic review expressed less trust in *"drug companies"*, *"insurance companies"* (Mello *et al.*, 2018, p. 2205), "*third parties"* (Jao *et al.*, 2015a, p. 10; Manhas *et al.*, 2015, p. 92), *"industry-based researchers*" (Manhas *et al.*, 2015, p. 94) and *"for-profit"* research groups (Manhas *et al.*, 2015) than in scientists and universities. They preferred to keep data in the *"research eco-system"* where possible (Mozersky *et al.*, 2020). Although the questionnaire respondents and the participants represented in the systematic reviews were less likely to be willing to have their data shared with pharmaceutical companies, they did not reject the idea outright. Previous research into attitudes towards sharing health records and biological data has also identified that participants prefer their data to be used in academic research than pharmaceutical or for-profit research (Trinidad *et al.*, 2010; Hendrix *et al.*, 2013; Shabani *et al.*, 2014; Aitken *et al.*, 2016b; Garrison *et al.*, 2016). Some participants are concerned about *"slippery slopes"* with data being shared further and further from originally agreed projects as time goes on (Aitken *et al.*, 2016a, p. 16).

Not mentioned in the systematic review was sharing of data with the government; however, just over a third of questionnaire respondents said that they would share with the

government to help them 'study health problems'. A similar percentage of questionnaire respondents (34%) said that they would be happy to share their data for student projects. Only two papers (Hate *et al.*, 2015; Mozersky *et al.*, 2020) included in the systematic review asked participants about student access but identified mixed responses. Mozersky *et al's* participants would be happy for students to have access to data for training in analysis techniques but Hate *et al's* respondents were split, with some being happy to share with students and others suggesting that students should make the effort to collect primary data for their own education (no numbers given as these were qualitative studies) (Hate *et al.*, 2015; Mozersky *et al.*, 2020). Some papers in the systematic review did something that the questionnaire did not by discussing separately secondary researchers and their secondary projects. In terms of secondary projects, participants in the review were more likely to be happy for their data to be shared with projects that contributed to science (Mello *et al.*, 2018) or were similar in scope to the original research project (Mozersky *et al.*, 2020) and were less enthused by secondary research that was profit driven.

The grey literature provided no explicit mention of whether participants should (be asked to) approve of secondary researchers or projects, although this could be inferred from suggestions that they have the opportunity to re-consent which is covered below in section 6.5.4. Otherwise, it appears that participant input at the level of each individual secondary study has not been considered before.

It should be noted, in light of participants' preferences, that public interest as referred to in the grey literature did not exclude commercial organisations (Lowrance, 2002; Medical Research Council, 2011; UK Research and Innovation, 2018).

Benefits of and motivators to share

The questionnaire asked respondents which statements of potential benefit made them feel more positive about data sharing (Q9) and then presented a separate set of statements about which might motivate them to share (Q10). All potential benefits and motivations were selected by respondents, but the aspect of sharing that respondents found most beneficial was the potential to impact the research participants, i.e., the opportunity for rarer diseases and conditions to be studied more easily using combined data sets. The motivators to share most likely to be selected by respondents were assured anonymity and the chance to help others by contributing to research. The importance of assured anonymity (or privacy) has already been discussed above as a lack of anonymity was also identified by

participants in the systematic review and respondents to the survey as a main harm from sharing.

The systematic review participants also expected benefits to participants, including to the communities who contributed the data (Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015b; Mello *et al.*, 2018) and to the general public (Cheah *et al.*, 2015; Manhas *et al.*, 2015; Manhas *et al.*, 2016). Many of the systematic review papers were based in low-income country community settings for whom a direct benefit from sharing may be more immediately useful. In fact, the systematic review participants could also appreciate the benefits of sharing to science and discovery (Jao *et al.*, 2015b; Colombo *et al.*, 2019; Mozersky *et al.*, 2020) just as well as those responding to the questionnaire, but these were covered with less frequency and depth than the benefits to the participants themselves.

The potential benefits of sharing were not extracted from the grey literature as they are not strictly guidance; rather, they provide context. The benefits of sharing from the perspective of researchers are outlined in the background chapter of this study (Chapter 1 Background and Introduction, section 1.4). However, although not extracted, the benefits of sharing are implicit within the guidance and instructions given, for example when Castell *et al* suggest the composition of access committees, they are explaining that by using their approach, researchers can ensure that access decisions are transparent and standardised (Castell *et al.*, 2018). Benefits of sharing most commonly referred to by funders, journals and academics are increased research efficiencies, increased transparency, faster benefits to science and medicine, less burden on participants and maximising the use of participant contributions (Vickers, 2006; Walport and Brest, 2011; Ross and Krumholz, 2013; Carr and Littler, 2015). Some systematic review participants also mentioned the benefits such as efficiencies and better use of resources, but this was overshadowed by the 'bigger picture' benefits; those to participants themselves or to science.

Overall, these findings on the benefits of research data sharing suggest that participants are more likely to appreciate tangible, practical results from data sharing over theoretical ones such as more transparent science, and all the better if these practical benefits of sharing also directly benefit the participants themselves. The grey literature does not provide any motivators that might encourage participants to share, as from the perspective of the guidance documents, it is researchers, not the participants who are to be persuaded or encouraged into sharing. It might be useful for future guidance to outline the aspects of

sharing that participants' find most beneficial, so that researchers can communicate these to research participants during the consent process.

Which items are participants willing to share?

The questionnaire asked respondents which anonymised details about themselves they would be willing to share (Q11). The majority of respondents (between 41 and 53 percent) were 'very willing' to share all fifteen potential details. The systematic review contained a handful of references to which data specifically participants would or would not be willing to share. One paper (Mursaleen *et al.*, 2017a) found that Parkinson's patients thought data that should never be shared were those that could identify them or potentially influence third parties such as employers or insurers, whilst demographic details such as date of diagnosis or employment could be shared occasionally, and treatments or symptoms could always be shared. Inability to get insurance because of sharing data of a pre-existing condition or uncovering of illegal activity was mentioned in a second study (Mozersky *et al.*, 2020, p. 19). When participants mention such issues, it seems that they may have momentarily forgotten that the data to be shared would be anonymised, and that they should not be identifiable to insurers (or anyone else). The questionnaire respondents seemed to be just as willing to share details about illegal activity (illegal drug use) as they were any other health or demographic information. In fact, more respondents were 'very willing' to share data about potentially stigmatising use of illegal drug use than details of employment. Only 'mental health' achieved a lower endorsement of respondents 'very willing' to share details than employment did. Respondents who rated their health as 'average' or 'very poor' were significantly less likely than those who rated their health as 'very good' to be willing to share some details about themselves (mental health, HIV status, and height and weight).

In the systematic review 85% of participants with diabetes (Shah *et al.*, 2018) were 'happy' or 'very happy' to share details of medical history, genetic information, blood test results and lifestyle information. This is much higher than the percentage of questionnaire respondents who were 'very willing' to share details of diseases and conditions, medications or lifestyle information such as smoking, alcohol and illegal drug use (44.7-50.2%).

Some participants noted that if the data was anonymised, they had no concerns about sharing data such as blood results (Cheah *et al.*, 2018, p. 5), while others thought that they would not mind if the data were *not* anonymised if it was something like a diagnosis of but would want data anonymised if it referred to sexual activity or alcohol consumption

(Mozersky *et al.*, 2020, p. 19).

The fact that the number of respondents who were willing to share details about themselves was very similar for each type of data (with a range of just 10%) could mean that respondents were not fully engaged with these questions. Giving 15 options (types of data) for respondents to consider could mean that the majority selected 'very willing' for most sub questions without really considering each type of data as being distinct from the others, or perhaps, like some of the systematic review participants, they were genuinely unconcerned as they knew the data would be anonymised.

Little reference is made in the grey literature to the types of data that could be shared, other than to stipulate that personal data or identifiers, both direct and indirect, are removed or anonymised (Lowrance, 2002; Corti *et al.*, 2014; Tudur-Smith *et al.*, 2015). The UK Data Service remind us that variables that may appear innocuous can in fact be indirect identifiers, such as age, workplace, occupation or location (UK Data Service, 2016), especially when presented in combination. References were also made to 'employment details' (which questionnaire respondents were cautious about sharing) in the grey literature. Lowrance referred to *"employer"* and Tudur-Smith *et al* and Corti *et al* referred to *"occupation or place of work"* (Lowrance, 2002, p. 42; Corti *et al.*, 2014, p. 119; Tudur-Smith *et al.*, 2015, p. 23). It would be interesting to know whether respondents were more wary of sharing employment details because they knew that occupation or employer could act as an indirect identifier (when combined with other variables) or whether there was another reason for this (e.g., concerns about the employer being able to access responses). It would also be useful to define whether respondents were most concerned about sharing employer or occupation as it could be argued that employer is potentially more identifiable that occupation and that occupation is of more use to researchers (for defining socio-economic status).

What we can perhaps conclude from all sources is that sensitive data items include, employment, mental health, gender, alcohol, illegal activity, and of course personal information, but all of this is context specific.

6.5.2 Does knowing about research data sharing affect likelihood of participation in primary research?

There was just one question in the questionnaire regarding whether respondents knowing their data might be shared in the future would influence their likelihood of agreeing to take

part in the original, primary research study. Most respondents said that, hypothetically, knowing that their data might be shared would have no effect on their taking part in (primary) research whilst some said that they would be a bit more cautious about taking part. In fact, respondents who had a positive or neutral experience of taking part in research already were significantly less likely to take part in a future hypothetical study than those with a very positive experience, after being informed that their data might be shared.

Participants in the systematic review studies exhibited mixed degrees of understanding of data sharing, with some being unaware whether or not their data might already have been shared (Asai *et al.*, 2002; Hate *et al.*, 2015; Jao *et al.*, 2015a; Jao *et al.*, 2015b; Manhas *et al.*, 2015; Manhas *et al.*, 2016; Mursaleen *et al.*, 2017b). Once participants were informed about data sharing, they generally wanted more information about it, including with whom their data might be shared, but overall researchers were trusted to make sharing decisions on participants' behalf (Manhas *et al.*, 2015; Mozersky *et al.*, 2020). With the exception of one participant who stated that they would want to opt out if their data were to be shared with a profit driven research group (Manhas *et al.*, 2015, p. 94), the systematic review did not identify any evidence exploring whether or not knowing about sharing would discourage participants from taking part in the first place. This question was not specifically asked of participants in the systematic review.

The grey literature also said little about how knowledge of data sharing might influence decisions on taking part in research, although it was acknowledged that participants could not give their informed consent to share and would not know what they were signing up for if the likely recipients of shared data were not disclosed during the consent process (Lowrance, 2002; The Academy of Medical Sciences, 2013; Corti *et al.*, 2014). Consent is discussed in section 6.5.4 below.

6.5.3 What are the preferences of research participants for data sharing?

Questionnaire respondents were asked about their preferences for consent, storage of and access to their data, as well as feedback on use of data and finally data ownership. The systematic review and grey literature review also provided evidence on these sharing processes.

6.5.4 Consent

The majority of questionnaire respondents (66.3%) thought it 'very important' that the consent form lets participants know that their data will be shared. This aligns with the evidence from the systematic review where, in one study, approximately 55% of participants thought that their consent should be sought before data was anonymised and prepared for sharing (Manhas *et al.*, 2018). Other evidence from the systematic review supported this, with participants preferring to be informed about the potential for sharing, at least as a courtesy (Mozersky *et al.*, 2020) and suggesting that the consent form could have an educational role in explaining data sharing to participants (Merson *et al.*, 2015). Nevertheless, just because participants are generally happy to share, it should not be assumed that researchers should do so without asking. Without an understanding of what data sharing involves or in the absence of appropriate information being provided during consent for data sharing, researchers can end up with *"at worst… an unsafe consent"* (Mursaleen *et al.*, 2017b, p. 527). The grey literature concurs that the consent process should let participants know that their data might be shared, with some also suggesting provision of details of with whom or for what purpose data might be shared so that consent can be *truly* informed (Lowrance, 2002; The Academy of Medical Sciences, 2013; Corti *et al.*, 2014). Informed consent is therefore a three-step process; first, participants must understand the information given about data sharing, second, they should be provided with potential future uses of data, and third, use this information to give informed consent.

<u>Type of consent</u>

Consistent with some other research (Hoeyer *et al.*, 2004, p. 227; Ludman *et al.*, 2010; McGuire *et al.*, 2011; Chan *et al.*, 2012; Clerkin *et al.*, 2013; Taylor and Taylor, 2014), the questionnaire respondents exhibited a slight preference for re-consent each time their data are shared, with giving a one-off consent at the beginning of the original study a close second.

The systematic review provided a great deal of data regarding participants' preferences on consent type, with some evidence suggesting that participants would prefer to engage little with researchers once the original research was over, avoiding the burden of re-contact for re-consent each time their data was requested for sharing (Jao *et al.*, 2015a; Merson *et al.*, 2015; Manhas *et al.*, 2016; Cheah *et al.*, 2018; Manhas *et al.*, 2018; Mello *et al.*, 2018). This re-contact burden was also recognised in the grey literature (The Academy of Medical

Sciences, 2013). There were fewer reports in the systematic review of participants preferring to consent each time their data was requested or shared (Asai *et al.*, 2002; Mursaleen *et al.*, 2017b; Cheah *et al.*, 2018).

However, if we look at questionnaire responses from individual studies, only ALSPAC respondents exhibited a preference for consent each time data was shared whilst ACONF and the PPI groups had majorities of respondents preferring one broad consent. Just 4.9% stated that they were happy for their data to be shared without being consulted at all and these individuals were significantly more likely to be older respondents or those in poor health. There was insufficient breakdown by participant health status and age in the systematic review to confirm whether this finding is common to all participants who are older or in poor health. The difference between the attitudes towards consent types in the questionnaire respondents and systematic review participants could be for any number of reasons, but it is possible that many of the systematic review participants were able to talk through consent types with researchers in the context of a qualitative interview or focus group, before concluding that they preferred not to be re-contacted each time for practical or nuisance reasons.

The grey literature, like the systematic review papers, discussed pros and cons of the various consent types, which are described variously by different organisations but referred to here as broad one-time consent or re-consent each time data are shared. In the grey literature guidance documents, these two consent types are largely discussed in terms of convenience to researchers, rather than preference of participants. For example, a re-consent model where participants are re-contacted to approve or decline a sharing request is burdensome for researchers in terms of cost and time and can result in a biased sample if only participants who opt-in or can be contacted are included in a shared data set (Lowrance, 2002; Medical Research Council, 2017). As exhibited in the systematic review, this repeated re-contact would be potentially burdensome for the participants too, and both parties would have to ensure that contact details (and preferences) were up to date. A broad consent model, where participants give a one-off consent to all future sharing during the consent to the original study (Medical Research Council, 2017) is much less burdensome for researchers but can still result in bias; if participants who do not consent to future sharing are removed from shared datasets, these datasets will never match the data analysed for the original study.

Information that the consent form should include

Questionnaire respondents were asked what information they would like to see on the consent form. The most popular items for inclusion on the consent form were explanations about data sharing, how their identity would be protected via anonymisation, with whom the data might be shared and who might benefit from this sharing. What systematic review participants wanted to see on the consent form was said to vary by participant group (Cheah *et al.*, 2018) but the evidence from all included studies can be summarised as: acknowledgment that their data may be shared and explanations and assurances of anonymisation; these findings support the preferences of the questionnaire respondents. As discussed above in section 6.5.1, systematic review evidence shows that participants were certainly also interested in with whom the data might be shared.

The grey literature provided guidance for researchers about information that should be provided on the consent form (Lowrance, 2002; Corti *et al.*, 2014; Medical Research Council, 2017; Castell *et al.*, 2018) and this was primarily assurances about anonymisation and a suggestion of how data will be used whilst avoiding *"information overload"* or impenetrable language (Castell *et al.*, 2018, p. 44). The grey literature was one step ahead of the systematic review and questionnaire as it was working on the assumption that data sharing definitely would be mentioned on the consent form. The questionnaire and the systematic review allowed for the possibility that participants may not be aware or informed at all. This could well be the case for participants who have previously consented to older studies with no mention of sharing on the consent form and may therefore be unaware their data is being or could be shared. Castell *et al* go further than some other guidance documents and suggest that participants should also be told who will access the data, how research findings will be shared, how access committees work, who sits on them, potential risks of sharing and how participants' data will be used in research (Castell *et al.*, 2018). The MRC also suggest researchers explain long term plans for sharing, archiving and publishing and who will be responsible for keeping data safe (Medical Research Council, 2017). The questionnaire respondents and systematic review participants seemed less interested in storage or access issues as compared to potential harms; therefore, following the recommendations of Castell *et al* and the MRC could in fact contribute to information overload. By contrast, Corti *et al* and The Academy of Medical Sciences place emphasis on confidentiality via controlled access; in these circumstances, it might be less important to provide specific examples of

what the data might be used for as control can be applied at the application stage (The Academy of Medical Sciences, 2013; Corti *et al.*, 2014). This approach, with data held in a repository that has appropriate security controls, might also provide some reassurance for participants, as a primary concern in both the systematic review and the questionnaire survey was data being stolen. Nonetheless, it is not known whether, when future uses are ill-defined or unknown, participants are happy with the 'unknown' as long as they are reassured that their data are securely stored and anonymised.

Evidence only apparent in the grey literature

Some aspects of research data sharing covered in the grey literature were not identified in the systematic review or subsequently explored in the questionnaire. Withdrawal of consent is considered by only two guidance documents (Medical Research Council, 2017; Castell *et al.*, 2018) but hardly at all by the systematic review participants. The MRC expect that participants are given the option to withdraw from a study and set out the different degrees of withdrawal available, including withdrawing from future analysis, although it is not clear whether this encompasses secondary analysis (Medical Research Council, 2017). Castell *et al* reported that the option to withdraw would give participants a sense of control over use of their data (Castell *et al.*, 2018). The MRC point out that there is a point beyond which participants will not be able to withdraw, and that this should be indicated on the consent form (Medical Research Council, 2017). The MRC give the example of post publication as a time at which participants could no longer withdraw, but it would also be difficult to withdraw a participant once their data has been anonymised and shared.

Another aspect of sharing not explored in the systematic review was sharing of data in circumstances where consent had not been obtained in advance, for example from studies where data were collected before mentioning this in the consent form became the norm. Guidance documents from several sources (Lowrance, 2002; The Academy of Medical Sciences, 2013; ESRC, 2015; Tudur-Smith *et al.*, 2015; Medical Research Council, 2017) maintains that sharing should not be prohibited just because consent was not sought as part of the original study, with advancement of science (secondary research) taking precedence over participant concerns such as (re)identification. However, Lowrance and the MRC also suggested that, where practicable, consent could be sought if the participants were still contactable and it would not be too costly or impractical, and that if this was not possible or practical, researchers must ensure that the data are anonymised (Lowrance, 2002; Medical

Research Council, 2017). This guidance does not acknowledge that some participants regard consent as a courtesy or a principle (Manhas *et al.*, 2015; Mozersky *et al.*, 2020). What is acknowledged, however is that there is a risk of the introduction of bias in the data included for analysis when researchers are only able to include data of participants who were re-contactable and (re-)consented (Lowrance, 2002).

6.5.5 Storage and access

Storage and access types

A large majority of questionnaire respondents (86.8%) preferred that their data was stored with controlled access, with only 3.7% preferring open access. By contrast, some systematic review participants from one study (39%) said that they believed access should be *"broad"*, open not only to researchers, but also *"other groups and individuals such as patients' and citizen group representatives and journalists"* (Colombo *et al.*, 2019, p. 5). Most grey literature aligned with participants' and respondents' views and advised caution in the approach to access, with only Research Councils UK advocating open access and as few restrictions as possible (albeit responsibly) (UK Research and Innovation, 2018). Crucially, no clinical trials units surveyed and referred to in Tudur-Smith *et al's* guidance advocated for an open access model (Hopkins *et al.*, 2016) cited in (Tudur-Smith *et al.*, 2015, p. 10).

The grey literature discussed pros and cons of open access versus controlled access, and although the greater risk of identification of participants in an open-access model is highlighted, researcher concerns such as the difficulty in tracking publications and research arising from the data are also identified (Tudur-Smith *et al.*, 2015). The Academy of Medical Sciences (The Academy of Medical Sciences, 2013) also identify the potential for reputational risk to the original researchers if participants were to discover that their data had been shared with commercial companies. This echoes questionnaire respondent lesser concerns about their data being used to make a profit or used in research they did not approve of and the views of systematic review participants who also wanted their data to be used in projects that align with their values (Mello *et al.*, 2018; Colombo *et al.*, 2019; Mozersky *et al.*, 2020).

The grey literature referred to storage both as a place to keep the data and a method of advertising the availability of the data; data needed to be discoverable and accessible in accordance with the FAIR principles (Wilkinson *et al.*, 2016) for *"maximum exploitation"*

(ESRC, 2015, p. 3). There is also some suggestion that storage options need to be considered at the outset of a study, not just because data will need to be stored for a long period of time (to comply with funder stipulations and to maximise sharing opportunities), but also because participants should be advised for how long their data will be stored (Medical Research Council, 2017). Where data are stored will influence how long it is stored for, as this may vary by repository, and could even mean data are available indefinitely. Long-term storage will mean long-term security and access decisions to be made. To honour participants' consent, these decisions will need to be consistent over the lifespan of the data set.

Privacy and security

For participants in the systematic review, privacy, security and storage were intertwined, with anonymisation providing reassurance of privacy, regardless of whether or not participants fully understood the actual anonymisation process (Mursaleen *et al.*, 2017a; Mursaleen *et al.*, 2017b; Cheah *et al.*, 2018; Shah *et al.*, 2018; Mozersky *et al.*, 2020). Anonymisation was not covered in great detail in the questionnaire, but we know from the responses to question 10 that respondents found it one of the single most important motivators for sharing. Systematic review participants also wanted assurances that their data were secure and that processes were in place to provide this security. This was particularly true if the data were sensitive in nature, such as that concerning their children or potentially embarrassing health conditions (Mozersky *et al.*, 2020). In some of the grey literature, security and anonymisation were discussed together as methods of protecting the privacy of research participants rather than referring entirely separately to secure storage of data and privacy through anonymisation (The Academy of Medical Sciences, 2013; Corti *et al.*, 2014; ESRC, 2015). It is not stated specifically, but security can be achieved by choosing secondary projects carefully and ensuring that no inappropriate or unapproved data linkage takes place. Lowrance suggests that researchers emphasise to participants that in anonymised data they are interested in *"cases"* not *"persons"* to provide reassurance of privacy (Lowrance, 2002, p. 27). For data that are less likely to enable re-identification, such as that which is highly aggregated, some guidance suggests that access controls can be less stringent (The Academy of Medical Sciences, 2013; ESRC, 2015).

Where access needs to be *more* stringent, security is not just about anonymising, but ensuring access controls. The ESRC suggest that variable access levels could be applied

accordingly depending on the sensitivity of the data (ESRC, 2015) and some organisations (The Academy of Medical Sciences, 2013; Corti *et al.*, 2014; ESRC, 2015; Tudur-Smith *et al.*, 2015) suggest that researchers accessing data for secondary research do so through use of a *"secure access infrastructure"* (ESRC, 2015, p. 5) or a *"protected virtual environment"* (Corti *et al.*, 2014) so that the data do not need to leave the site or network of the original research organisation. Instead, secondary researchers select the required data and perform analysis without receiving the data themselves. Allowing secondary access through the original research team would also align with participants' preferences since, although participants had trust in researchers in general, trust in the original research team was greater than that in unknown secondary researchers (Manhas *et al.*, 2015; Manhas *et al.*, 2016; Mello *et al.*, 2018; Mozersky *et al.*, 2020). In this aspect, the guidance aligns with concerns of participants.

Access decisions

When it came to making decisions about granting access to stored study data, if data had controlled access, about 38.8% of questionnaire respondents thought that the participants who took part in the study should give permission for their data to be shared, with the next popular answer 'the organisation where the data was collected' (23.3%). Some participants in the systematic review thought that governance issues were more important than privacy (Manhas *et al.*, 2018), and there were many references to committees or gatekeepers who could make decisions on access or sharing requests (Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Merson *et al.*, 2015; Manhas *et al.*, 2016; Cheah *et al.*, 2018). Ideally committees or a *"group trusted to make decisions"* (Jao *et al.*, 2015a, p. 271) would contain lay members (Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015) or experts (Shah *et al.*, 2018) who could reach a consensus. There was also an interesting mention of committees being accountable for their decisions (Manhas *et al.*, 2015) with penalties for data breaches or misuse (Manhas *et al.*, 2015; Colombo *et al.*, 2019), showing the strength of participant feeling.

By contrast, few questionnaire respondents chose a committee as their preferred option for making sharing decisions (9.2%). In agreement with systematic review participants, the grey literature also advocated the use of independent committees which could make transparent decisions about sharing regardless of their membership composition. The Academy of Medical Sciences talk about independent panels as a new way, acting as both data holder

and as a decision maker (The Academy of Medical Sciences, 2013, p. 3). Committees could be made up of data custodians, (e.g.: clinical trials units) with further advice sought from sponsor and an independent committee when required (Tudur-Smith *et al.*, 2015), or a group or individual that is separate from the data custodian (The Academy of Medical Sciences, 2013). Just one guidance document mentioned the possibility of lay membership of access committees (Castell *et al.*, 2018). The participants who informed Castell *et al's* guidance, like those in the systematic review, were insistent on non-tokenistic lay membership of access committees. It is unlikely that lay membership of access committees would have been identified in the grey literature had Castell *et al's* report not been influenced by stakeholder input, including from research participants. Use of committees could remove the need for the original researchers (who may be transient) to remain involved in the project so that they can make decisions about whether the data can be shared.

Despite this being their preference, the questionnaire was not able to address with respondents how they might be able to make access decisions themselves, but presumably they would be happy with further contact from researchers so that they could agree to or opt out of secondary research. Alternatively, the original consent process could seek agreement as to the types of sharing that would be acceptable. It should be noted that the majority of questionnaire respondents were from ALSPAC or ACONF and were therefore familiar with re-contact by their original study team regarding further research and therefore envisaged that all research participants could be contacted in this way. Similarly, some of the systematic review participants reported that they trusted the original researchers to make sharing decisions on their behalf (Manhas *et al.*, 2015; Manhas *et al.*, 2016; Mello *et al.*, 2018; Mozersky *et al.*, 2020). Researcher-specific processes and procedures identified in the grey literature but (unsurprisingly) not discussed in the systematic review with participants include the use of data management plans and data access agreements and provision of supporting metadata or documentation alongside shared datasets.

6.5.6 Ownership

When asked about ownership, the greatest number of questionnaire respondents thought that 'the participants who took part' (49.3%), 'the researcher(s) who collected it' (45.2%), closely followed by 'the organisation where the original researcher(s) work' (44.1%) owned

study data. In the systematic review only two papers but one researcher (Mursaleen *et al.*, 2017a; Mursaleen *et al.*, 2017b) reported on ownership; in that work, more than a third (38%) of focus group participants thought that data were owned by whoever it had been shared with, 24.1% thought that the participants themselves owned it, and 14.5% of participants thought that the platform upon which it was shared could claim ownership. Eighteen percent of participants simply did not know (Mursaleen *et al.*, 2017b). This is similar to the questionnaire respondents in that there are mixed responses, but it seems that Mursaleen *et al's* participants thought that data are more likely to be owned by the recipient or user, whereas the questionnaire respondents thought that data are owned by those who collected it or the participants themselves. It makes sense that if respondents believe that they own their data they should also make their own sharing decisions (Q20). This was also exhibited by Mursaleen *et al's* participants who thought that to share the data, one should own it (Mursaleen *et al.*, 2017a). Mursaleen *et al* therefore advised that any confusion about data ownership is avoided by clarifying as part of consent who would own data and make sharing decisions (Mursaleen *et al.*, 2017b).

Most of the documents included in the grey literature review do not mention or identify a definitive data owner, and ownership of data is mentioned infrequently. The UKRI call for *"clarity"* on ownership of research data, but this seems to be more for the benefit of researchers than for participants (UK Research and Innovation, 2018). Tudur-Smith *et al* provide this clarification, reminding us that clinical trials units are just data *"custodians"* and that sponsors of studies are the real data owners who need to be involved in any decisions about sharing (Tudur-Smith *et al.*, 2015). By contrast, the ESRC state that, unless agreed otherwise, the organisation conducting ESRC funded research becomes the data owner, with responsibilities to exploit the data for benefit (ESRC, 2015). Two of the guidance documents included in the grey literature review (Corti *et al.*, 2014; ESRC, 2015) briefly referred to ownership in the context of recognition of the contribution of the original research team. But this is not the context in which participants in the systematic review discussed ownership, and neither was that the intended context of the question about ownership in the questionnaire.

Lowrance explains that ownership of data is actually unrelated to use of data because although participants do not own their data (contrary to the perceptions of many questionnaire respondents and systematic review participants), they still must give

permission for it to be used (Lowrance, 2002). So, the issue is not of ownership, it is of permission or consent. All of this needs to be unpicked and explained to participants as part of the consent process.

6.5.7 Feedback

A large majority of questionnaire respondents (81.9%) thought that feedback on how their data was used was 'important'.

Five papers in the systematic review referred to feedback to participants' regarding use of their data for secondary research (Asai *et al.*, 2002; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Mursaleen *et al.*, 2017b; Manhas *et al.*, 2018), but only two really give detail on participants' preferences. The majority of Manhas *et al's* participants wanted to hear from researchers once a year (Manhas *et al.*, 2018), whilst the majority of Mursaleen *et al's* wanted to be informed by email when their data were used (Mursaleen *et al.*, 2017b). Manhas *et al* also suggested a password protected account for participants to provide them with information on how their data were being used (Manhas *et al.*, 2018) which could provide a way of informing participants without over-contacting them and could save time for researchers. Participants in the study by Manhas *et al* (2018) preferred any communication to be about projects in which they took part as opposed to other datasets, which echoes what other participants said about the burden of re-contact (Manhas *et al.*, 2018). Shah *et al* refer to GDPR and suggest that, to keep consent valid and provide the required opportunities for withdrawal, participants would need to be contacted each time their data are used (Shah *et al.*, 2018), which would please those questionnaire respondents who expressed a preference for re-consent.

Some of the guidance documents in the grey literature review refer to provision of feedback to participants about use of their data in secondary research, and that which most closely echoed participants or respondents' concerns was Castell *et al* (Castell *et al.*, 2018); this is not surprising since this guidance document was informed by consulting research participants. Like the systematic review participants, the participants in Castell *et al* were interested in feedback on how and how many times their data had been used but recognised the practical difficulties of doing so. Providing blanket feedback also ran the risk of bothering participants who had not requested any feedback and so Castell *et al* suggested that a summary of data use might be presented on a website, echoing Manhas *et al* (2018). Other guidance stated that feedback methods should be considered at the outset of a study

(Medical Research Council, 2017) and that feedback was important where feasible, although could risk bothering participants who had not wished to be re-contacted (The Academy of Medical Sciences, 2013).

6.5.8 Summary of triangulation

There was a large degree of corroboration between the systematic review data and the respondents' answers in the current survey. Questionnaire respondents and systematic review participants were generally open to research data sharing, albeit with greater caution about sharing with pharmaceutical companies. Questionnaire respondents wanted more control over sharing of their data than their systematic review counterparts, with respondents expressing a slight preference for consenting each time it was shared, a preference not exhibited by most of the systematic review participants. Both groups were concerned about privacy. The grey literature review presented a lot of information on consent types; along with access to data and anonymisation, this was probably the aspect of data sharing that attracted the most guidance, perhaps because these matters are something tangible that can be tackled by researchers. Some guidance suggested the provision of examples of how data might be shared in advance, and some even suggested that re-consent should be sought where possible.

6.6 Strengths and limitations of the PhD study

6.6.1 Strengths

There are a number of strengths to this PhD study which are outlined below under the relevant sub-headings.

Triangulation

This study incorporated two literature reviews; a systematic review of international literature on participants' attitudes and a grey literature review of UK guidance, meaning I can be confident that as much relevant published evidence as possible has been included. This evidence was then combined with the results of the questionnaire survey before 'triangulating' the results. So that the volume of evidence was not overwhelming or unwieldy, data from the grey literature review and questionnaire survey was gathered using a key theme of the systematic review, which was conducted first. This theme 'conditions and pre-requisites' identified that review participants were most concerned or interested in consent for sharing, storage, and access of data and with whom, or for what purpose that

data would be shared. These topics were then used to extract evidence from the grey literature and to guide the flow of the questionnaire survey. By following this theme throughout the PhD study, the evidence can be more usefully compared at the end in this discussion chapter. This also means that the recommendations made, based on this evidence, will be responding to real participant concerns.

Systematic review

One of the key strengths of the systematic review was the extensive search of the international literature, which increased the likelihood of capturing all relevant published evidence. Both qualitative and quantitative studies were eligible for inclusion resulting in a total of eighteen included studies. The published version of the systematic review (Howe *et al.*, 2018) was, as far as I know, the first review to explore attitudes towards secondary use of trial or health study data.

The systematic review methods including the quality appraisal, data extraction and thematic synthesis were made explicit so that they could be reproduced. All of the included studies were found to be of reasonable or high quality, providing support for the validity of the results. Included studies originated from a wide range of settings and countries including those that were high, middle- and low-income, potentially enhancing generalisability of findings. The systematic review was also updated relatively recently (August 2020) to ensure that all relevant papers are included.

Grey literature review

The grey literature review focussed on guidance from UK organisations so that the advice given could be compared to the views of UK respondents who took part in the survey. A systematic approach was taken to the search for relevant literature and subsequent application of inclusion criteria. Methods were reported clearly in the review so that they can be reproduced. Data extraction was also conducted systematically, with data extracted under predetermined categories- the key concerns of participants as identified in the systematic review. This ensured that the extracted material was relevant not just to researchers, but to participants and allowed triangulation of findings.

Questionnaire survey

This questionnaire survey achieved a relatively large sample (1,664 respondents); this sample size is larger than most other published surveys on attitudes to data sharing, both

those included in the systematic review (Mursaleen *et al.*, 2017b; Manhas *et al.*, 2018; Mello *et al.*, 2018; Shah *et al.*, 2018; Colombo *et al.*, 2019) and those not eligible for inclusion (Willison *et al.*, 2009; Ludman *et al.*, 2010). The participants of the two longitudinal studies from which survey respondents were sampled ranged in age from their mid-20s to their 70s and were from opposite ends of the UK. By virtue of being a birth cohort (ALSPAC) and a cross-sectional study based on year of birth (ACONF), these two groups can claim to be representative of the general population, at least of the areas from which the cohorts were assembled.

The overall response rate for the questionnaire survey was approximately 16%, which although low, still yielded an ample number of responses for analysis; response rates were higher for some sub-groups (notably SAIL/SUPER group and ACONF) (see Chapter 5, section 5.11). The survey was conducted, and this sample was achieved, in difficult circumstances; few study investigators had permission for secondary contact of their participants and fewer were willing to burden their participants with additional contact inviting them to complete a students' questionnaire survey. For those researchers who did agree, the original research study team were burdening participants with additional and unexpected contact, whereby they would complete a survey that had no relevance to the original study. In addition, the survey was distributed during the COVID-19 pandemic, when it is safe to say that the attention of a lot of researchers and participants may have been elsewhere.

The majority of respondents to the survey were derived from two large-scale longitudinal cohort studies. While there are many such regional and national cohorts, with some dating back to the 1940s (for example: Pearce *et al.*, 2009), surprisingly there is currently very little published literature which examines the attitudes of longitudinal study participants to data sharing. A cursory search of MEDLINE revealed only two papers (Audrey *et al.*, 2016; Manhas *et al.*, 2018), one of which was included in the systematic review for this study (Manhas *et al.*, 2018) and so the results of this study, when published, will provide valuable corroborative evidence for researchers working with longitudinal studies.

In a quote found post questionnaire production, Oppenheim states that *"not everyone realises that the design of a survey, besides requiring a certain amount of technical knowledge, is a prolonged and arduous intellectual exercise…"* (Oppenheim, 1992, p. 7). The development of the questionnaire survey was certainly an intensive process, involving scouring the existing literature to identify questionnaires that had already been used to

measure attitudes to sharing, conducting a scoping focus group to identify current UK based participant concerns and a process of questionnaire self-assessment, readability testing and cognitive interviewing. The questionnaire was as good as it could possibly be, given the time and resource constraints of a PhD project. A robust and well-designed survey as a measurement tool is more likely to answer the study questions, prevent questionnaire drop out, and to reliably measure attitudes (Oppenheim, 1992; Thwaites Bee and Murdoch-Eaton, 2016;Eaden *et al.*, 1999).

6.6.2 Limitations

However, there are a number of limitations to the PhD study, and it is important that these are acknowledged.

Triangulation process

There was some difficulty in comparing the results of the systematic review, which subjectively and qualitatively summarised data (sometimes from qualitative studies) in a narrative account, with data from the questionnaire which provided a quantitative output. Questions were not always asked in the same way, for example, some of the papers in the review were qualitative and more exploratory whilst the questionnaire captured the same data with a measurable scale. However, in most cases, both sources were providing similar evidence, meaning we can be more confident in the reliability of the results (Green & Thorogood, 2014). The systematic review and subsequently the questionnaire were not exhaustive and did not ask respondents about every aspect of data sharing. As outlined above in section 6.5.4, page 251, there are some aspects of sharing identified in the guidance that were not explored by me or by the researchers whose work appeared in the systematic review, and there may be more that remain unidentified. In addition, I deliberately asked respondents fewer questions about data storage and access, based upon feedback from participants who took part in cognitive interviewing and on the experience of asking the focus group participants about storage and access to data.

The data from the systematic review was also worldwide while the questionnaire was distributed in the UK only, and the guidance documents all related to UK practice. Varying levels of experience and expectations across countries and different cultural norms may underpin some discrepancies between the survey and systematic review findings.

Systematic review

One limitation of the systematic review is that six of the eighteen included studies (Cheah *et al.*, 2015; Hate *et al.*, 2015; Jao *et al.*, 2015a; Jao *et al.*, 2015b; Merson *et al.*, 2015; Cheah *et al.*, 2018) originated from the same research team and funder and were set in low and middle-income countries. Any similarities in findings may be due to comparable methodologies or populations. Most of the other studies were also from outside of the UK, and indeed there was only one study from Europe. Two studies by Manhas *et al*. used data from one research project (Manhas *et al.*, 2015; Manhas *et al.*, 2016), and a third included study was by the same author (Manhas *et al.*, 2018). Mursaleen *et al* contribute two papers about attitudes of Parkinson's patients to this review (Mursaleen *et al.*, 2017a; Mursaleen *et al.*, 2017b).

It is possible that participant quotes were taken out of context when subjectively coding the data during the thematic synthesis, although the original texts were referred to as often as possible. By interpreting sections of text as reported by other researchers, assigning codes and deducing themes it is possible that key participant concerns could have been missed entirely due to the influence of my own pre-conceptions or ideas.

Related to this, and as mentioned in the systematic review chapter (Chapter 2, section 2.5.2 fears and harms misuse), some of quotes from the included papers, and some of the explanation provided by the authors of the papers do not make it entirely clear that participants fully understood the process of sharing or anonymisation or were speaking about sharing of anonymised data when quoted. For example, it is difficult to imagine that harms such as identity theft, inability to gain insurance and attempted abduction of their child (Manhas *et al.*, 2015; Mozersky *et al.*, 2020) could arise from sharing of a fully anonymised dataset. Not all of the papers explained the extent to which participants were reminded about anonymisation when answering the researchers' questions, or whether, when questions were asked, a distinction had been made between sharing of anonymised data, pseudonymised data or fully identifiable data used in the study. Participants may, without reminders, conflate the pseudonymised data used to run a study with the anonymised data that is shared.

Grey literature review

The limitations of the grey literature review are discussed in detail in Chapter 3 (section 3,13), but briefly, the main limitations of the grey literature scoping review were the difficulty in identifying detailed and up to date guidance from the UK, specifically on data sharing. For some organisations, data sharing was just a part of overall data management activity whilst for others, data sharing guidance comprised of statements of intent without clear instructions of how researchers should implement them.

Some guidance documents (e.g. (Lowrance, 2002; Medical Research Council, 2016; NIHR, 2019)) did not always make the distinction between anonymised, pseudonymised or identifiable data which made interpreting the recommended best practice more difficult. There was some advice given to researchers regarding management of administrative data used for the purposes of running a study, separately from the resultant study dataset itself (Corti *et al.*, 2014; Medical Research Council, 2017), but again not all guidance made this distinction.

The decision to summarise the eligible guidance documents within the framework of systematic review topics could have led to key guidance that did not fall into the selected topic areas being omitted. If a concept or topic was not identified in the systematic review it was not searched for within or extracted from the grey literature. It is also possible that an entire informative document was excluded due to omissions from the search strategy or the inclusion criteria.

Finally, as with the systematic review, it is possible that extracting and summarising guidance documents has resulted in their messages being misrepresented or taken out of context.

Questionnaire

Questionnaire length and chance findings

One area where the questionnaire could have been improved was the length. The questionnaire was long, and some, but not many respondents seemed to get fatigued towards the end of the questionnaire, with fewer responses to later questions than to those earlier in the questionnaire. A long questionnaire with many dependent variables also means a greater number of cross-tabulations for secondary analyses and significant findings (this study had a total of 214). The more models, or in this case cross-tabulations including the Chi-squared statistic, that are run, the more likely a chance finding of statistical significance

is through *"dredging"* or *"fishing"* the data (Gelman and Loken, 2014). For example, at the 0.05 significance level, we will observe a significant result incorrectly (due to error or chance) every 1 in 20 times (Skelly, 2011). Of the six hundred and thirty-two cross-tabulations made (each independent variable x each dependent variable), 214 significant results with a Chi squared value that was significant at the 0.05 level were observed. To reduce the number of significant results that may have occurred by chance, a more conservative significance level can be selected (De Vaus, 2014, p. 229). However, even at the 0.025 significance level, 184 significant results were still observed for this study, predominantly related to the age of respondents, their experience of taking part in research and source study. In addition, this was an exploratory analysis and there were no hypotheses to be tested, so for clarity I decided to present results without making explicit further consideration of the potential for Type I error (observing a statistically significant association, when in fact no such relationship exists apart from in the particular data set) (Bobashev, 2011). I therefore decided not to implement a more conservative level of significance than the 0.05 previously stated. Instead, some caution was applied when interpreting the results, especially those that were not consistent with other evidence such as that from the systematic review (Savitz and Olshan, 1998; Gelman and Loken, 2014).

Bias

Perhaps a more significant limitation is that, in common with all survey research, the questionnaire element of the survey was open to inherent bias, both in respect to the sample of respondents selected and the responses to the questionnaire. This bias could serve to make the results of the questionnaire survey less generalisable. The specific types of bias affecting the survey results are explained below.

Sampling bias

Sampling bias occurs when the participants of a study are not chosen at random from the population and therefore there is a systematic difference between the achieved sample and the population to which the researcher wishes to generalise (Fink, 2003b; Berg, 2005; Flowerdew, 2005; De Vaus, 2014). The population of interest for this research was individuals who had personal experience of research participation or were interested in research. It was further intended that this study would focus more narrowly upon participants who had taken part in public health research, longitudinal studies, or clinical trials, meaning that the results could be generalised to these kinds of participants rather

than those who had shared other kinds of health data such as biological data or health records. Further to this, only studies which had ethical approval in place for participant re-contact, and where the investigators were willing to grant me access to their participants could be included. Those approached to take part in the survey ended up being participants in two longitudinal studies and members of patient and public involvement (PPI) groups, although the achieved sample was very much skewed towards inclusion of longitudinal study participants (96%). A further intended group had been parents whose child had taken part in a clinical trial, but ultimately these individuals could not be contacted due to administrative and COVID-19 restrictions. The sampled individuals and achieved sample are therefore not representative of the entire public health/longitudinal study/clinical trial population as it included no known clinical trial participants; nor can it be assumed that the views of ALSPAC and ACONF participants are generalisable to the underlying population of all longitudinal cohort study members.

Non-response bias

As well as sampling bias, there is the potential with any questionnaire survey for non-response bias. The survey was distributed by the administrators of the ACONF and ALSPAC studies to participants via email, therefore only participants who had provided their email address were contactable. For the ALSPAC study the questionnaire survey link was only sent to individuals who had previously stated that their preference was for online survey completion (n=5858). There were no reminder emails sent to participants who failed to respond to the initial emails and no incentives to take part due to budgetary restraints. It could also be argued that only the most altruistic of participants might respond to a request to take part in a research survey without a reward, particularly when those from ALSPAC are used to receiving a reward for their time.

Participants who chose not to respond to the survey invitation or did not see it were not represented in the survey respondents. Participants who prefer to use email to complete study follow ups may differ from those who chose to use paper, for example, perhaps they are more trusting of technology and would be more trusting of data sharing. Existing literature around questionnaire completion rates make it clear that, even within fairly homogenous groups, respondents who chose to take part in surveys may be systematically different from those who do not, thereby contributing to non-response bias (Fink, 2003b; Berg, 2005; De Vaus, 2014).

Comparisons of the characteristics of respondents from ACONF and ALSPAC as compared to the remainder of their cohort, and a comparison between all survey respondents and the general population were made in Chapter 5 (section 5.14). These comparisons identified that the achieved sample for ACONF was broadly similar to the underlying cohort with the exception of deprivation score, where those responding to the survey may be more affluent than the underlying cohort. From the ALSPAC cohort females and white respondents were over-represented but ALSPAC themselves have acknowledged that this is typical of their recent survey responses (Boyd *et al.*, 2013). When comparing respondents to the general population, it became apparent that again female respondents were over-represented but that the self-rated health of respondents is generally similar to that of the general population. There was some evidence of non-response bias in terms of ethnicity.

It is also plausible that those who chose to take part may be more interested in research in general, or in research data sharing than those who did not complete a questionnaire, and that responders may have stronger views (whether positive or negative) on the topic.

Response bias

Individual question non-response did not appear to be an issue for this questionnaire survey as all questions achieved more or less an equal proportion of responses (or missing responses). No attempt was made to impute missing data and instead respondents with missing values were excluded from analysis on a case-wise basis.

Of course, there is still the possibility of response bias or *"non-random deviation of the answers from their true value"* (Villar, 2011) in that respondents may have misunderstood questions, responded in a way that they felt reflected the most socially acceptable or desirable views, or provided the answers they think the researcher wants (Oppenheim, 1992).

Respondents' understanding of sharing

As with the systematic review participants, there is a possibility that the questionnaire respondents did not fully understand data sharing processes, for example, how shared data would be anonymised, or which data would be shared, although brief reminders were given in accompanying explanatory text. Questions or individual response categories may have been misinterpreted. Had respondents been able to ask for clarification on certain points whilst completing the questionnaire, their responses may have differed. Response bias

arising due to misinterpretation was countered as far as possible with a thorough survey development and testing (Chapter 4).

Type of data sharing

A final caveat of this research – albeit one that was embedded in the aims and objectives – is that it was focussed on just one type of research data sharing, i.e.: of data collected in the context of public health, longitudinal or clinical trials studies. There are other types of health data sharing (as identified in Chapter 1) such as sharing of routine health data (medical records) and biological data, data linkage and other types of *non-health* data sharing; had these been explored, different attitudes towards data sharing might have been revealed.

These limitations are all relatively minor and should not detract from the large volume of data and resultant evidence about research participant attitudes towards sharing that has been gathered.

6.6.3 Reflections on the PhD study

This PhD was conducted as a sequential mixed methods study where the qualitative work (scoping focus group and cognitive interviewing) informed the development of a questionnaire to collect data which were analysed quantitatively. My previous expertise has generally been in quantitative research; in my work as a database manager at Newcastle Clinical Trials unit, I am used to managing pseudonymised participant data and long-term projects, so this aspect of the PhD was not too difficult. However, the qualitative work was more nerve wracking, particularly conducting the focus group (although the participants put me at ease). I learnt from the qualitative aspects of this study how valuable the contributions of participants are; pointing out things I had not thought of, and confirming or contradicting the literature, and this emphasised to me that participants should drive policy that affects participants. This supports the inclusion of recommendations based on the evidence gathered from participants themselves.

My role means that I am cognisant of both the benefits and barriers to data sharing from the researcher perspective. However, my understanding of participants' attitudes towards sharing, and subsequent recommendations for best practice has been developed from my interpretation of the evidence in this study. My intention was to use these recommendations to protect participants' interests and ensure that data is shared ethically. It is possible, due to my position as one of the very researchers who often fail to consider

participants' views, that I have over compensated when it comes to my recommendations, being more cautious than necessary on behalf of participants, some of whom were not actually concerned about sharing.

6.7 Recommendations for future policy and practice

As discussed at the conclusion of the grey literature review (Chapter 3) few of the included guidance documents provide a comprehensive plan for data sharing from study set up to study close and beyond. Instead, researchers must search multiple sources. A further limitation, already noted, is that few of the guidance documents, and therefore recommendations available to researchers, have been informed by the views of potential or active research participants.

To fill this gap, and drawing on the evidence gathered for this study, from the grey literature, the systematic review and the questionnaire survey, a series of brief recommendations for researchers has been developed; these are presented below. It is not anticipated that these recommendations will require a great degree of policy change at a high level (e.g., from funders or journals), instead they comprise practical steps that researchers could make to align their research with the preferences of participants whilst still operating within the overarching guidance of their appropriate research funder. These recommendations are as follows:

6.7.1 Recommendation 1: Explain the rationale for sharing

To provide participants with enough information for their consent to future sharing to be informed, researchers should provide an introductory statement of what sharing actually is, including who might benefit and how data will be anonymised and stored securely prior to sharing as per the responses to Q13 of the questionnaire. McGuire *et al.* warn that a *"lack of specificity about data release in the informed consent process promotes variation in subjects' understanding and can lead to misunderstandings and false assumptions"* McGuire *et al*., 2008, p. 52). A fully informed consent has been linked both to trust and to a sense of control for participants (McGuire *et al*., 2008; Aitken *et al*., 2016a; Kalkman *et al*., 2019a; Broekstra *et al.*, 2020). Demonstrating security and privacy measures, explaining what research may be conducted and how studies could benefit patients has been said to increase trust and therefore the likelihood of consent for sharing (Damschroder *et al*., 2007).

The information provided should be understandable, 'framed' and 'selective' to be neither too detailed or too vague (Williams and Pigeot, 2017, p. 242) and participants should have the opportunity to ask questions about any aspects of sharing that they did not understand before signing consent. Researchers may also consider tailoring the information provided to the population who are consenting to sharing (Cheah *et al.*, 2018).

6.7.2 Recommendation 2: Explain which study data will be shared

Researchers need to explain to participants exactly which data will be shared. As explained above in the limitations (section 6.6.2), it is possible that participants still conflate sharing of anonymised research data with sharing of pseudonymised data analysed during the course of the original study, or with personal data (e.g., contact details) used in the conduct of the original study. Researchers should only share data that has been anonymised.

Researchers could incorporate some brief statements distinguishing between these types of data and who will have access to them, for example, that personal data will only be seen by clinical staff, pseudonymised data will only be seen by those running the study and that anonymised data is the only data that might be shared. Brief explanations of anonymised and pseudonymised (or synonyms thereof) will be required first.

If researchers are intending to contact (or allow contact of) participants about future research studies and are including a statement to the effect of 'you may be contacted about future research projects related to the original study', it would also be reassuring to participants to explain that although they may be contacted about future research, contact will be via the original research team and secondary researchers will never have access to their personal details.

6.7.3 Recommendation 3: To ensure fully informed consent, give examples of with whom or why data might be shared

A *sine qua non* is that those recruiting participants to a primary research study should include a statement about future research data sharing on their consent forms (a suitable statement for inclusion on consent forms can be copied from the HRA's website[9]). However, a simple statement about sharing is not sufficient.

[9] https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/data-protection-and-information-governance/gdpr-guidance/templates/template-wording-for-generic-information-document/

Researchers should then expand this to also include an additional simple statement regarding with whom data might be shared. Examples might include broad categories such as researchers (or students) in other institutions such as other universities, charities, hospitals (or other healthcare settings) or commercial organisations. This could be combined with a brief example of the purpose for which this data might be shared; 'further research' would likely suffice, but also possibly 'to combine with other datasets for analysis', 'cost-benefit analysis' or 'to inform research design'. The nature of the original study (and data) will constrain future uses, making them easier to predict. Bates *et al* link regulation with trust in research (Bates *et al*., 2010). By establishing ground rules with participants about acceptable types of projects, and then sticking to these rules when sharing, as well as implementing measures to ensure that there will be no deviations from these rules (or accepted projects) in any future sharing, researchers can foster trust in participants (Bates *et al*., 2010). If there is enough trust in the researchers running the study in the first place, broad categories such as these should be sufficient *"what makes it reasonable for study participants to invest trust on the basis of limited information and commitments about future actions"* (Williams and Pigeot, 2017, p. 248).

Researchers may wish to include a statement about with whom data would never be shared and whether data will stay within the same country. UKRI point out that *"all reasonable steps should be taken to ensure that research data are not held in any jurisdiction where the available legal safeguards provide lower levels of protection than are available in the UK"* (UK Research and Innovation, 2018, p. 17). By including this information, participants are not consenting to the unknown. It would be preferable for researchers to err towards being over-inclusive of potential future uses, to avoid having to re-consent or inadvertently deceiving participants by omission.

As with including a statement explaining data sharing, including a best estimate of with whom data might be shared on the consent form should be a pre-requisite for researchers. Participants want the aspect of control that knowing generally who their data will be shared with provides, but do not wish to *"micromanage"* by making sharing decisions themselves (McGuire *et al.*, 2008). If participants are unhappy with proposals for future sharing, they are able to decline consent for future sharing but agree to take part in the original research, as Haug explains, *"the patient shares, first with the clinical trialists and then, if the patient*

wishes, with data scientists" (Haug, 2017). Alternatively, the participant can make an informed decision not to take part in the original research study at all.

In the event that a study or researcher has no firm plans for data sharing at the stage of designing the original research study and recruiting participants but anticipate that they may accept sharing requests in the future, they should err on the side of caution and incorporate a statement about likely recipients as suggested above. If researchers go on to receive (and intend to approve) future sharing requests that fall outside of the examples given on the consent form, they should use re-consent at this stage. Accordingly, researchers should be careful not to promise *not* to share (Meyer *et al.*, 2018, p. 132) by stating that data will be *"kept private"* or *"only the research team will have access"*. Although data should absolutely be anonymised before sharing, researchers cannot promise that data will not be shared with, seen or managed by other researchers at some point in the future.

Researchers will also need to place the same conditions or limitations upon recipients of shared data to ensure that data is not irresponsibly re-shared or used in projects not originally consented to by participants. This can be achieved through use of data sharing agreements and/or sharing decisions which approve only projects which echo those that participants consented to.

6.7.4 Recommendation 4: Explain who will make sharing decisions

As well as providing information about who data might be shared with, researchers should provide a brief statement about who will make decisions on sharing on behalf of participants. It should be acknowledged that this could change over time, if and when the original researchers move on. Researchers should not bank on the trust that participants hold in the original research team (Manhas *et al.*, 2015; Aitken *et al.*, 2016a; Manhas *et al.*, 2016; Mello *et al.*, 2018; Mozersky *et al.*, 2020) to gain consent if they know that requests will be managed elsewhere. If a study has identified from the outset that data will be placed in a repository for sharing, this should be stated, along with a brief statement about the type of access (controlled or open) that will be used and whether decisions on sharing will be controlled by the repository, the original researchers themselves or a committee acting upon their behalf.

6.7.5 Recommendation 5: Use controlled access with independent review

Researchers should store study data with controlled access, that is by ensuring that data is not released to secondary researchers without some sort of request procedure being followed. In line with participant preferences, data should not be freely available ('on the internet' or in a repository) to just anyone. 'Governed' access is also recommended by most stakeholders (Bull *et al*., 2015) and by some of the literature included in the grey literature review (The Academy of Medical Sciences, 2013; Tudur-Smith *et al*., 2015). Access requests should then be evaluated by access committees or independent panels.

The IOM point out that controlled access can be viewed across a spectrum from requiring registration of requestees, through to checks of researcher qualifications and proposed analyses plans (Institute of Medicine (US), 2013); control needs only be as onerous as the data, researchers or participants require. This control is applicable both to data being shared in repositories and data being shared researcher to researcher.

Controlled access is not about preventing secondary research but about ensuring that it aligns with participants' desire to know who will access their data and with the consent that participants have given. Data requestors of whom participants would not approve or have not already consented to can be rejected or at least more carefully considered. In fact, the IoM have suggested that rather than completely blocking or refusing access to data, committees could provide advice to researchers who have been denied access on how to amend their application so that secondary research can still occur whist simultaneously protecting participants' preferences (Institute of Medicine (US), 2013). Controlled access does not refer just to the actual access to the data, but the controls applied to use of the data once access is granted i.e., those stipulated in data sharing agreements.

If researchers do decide to seek consent for complete open access, they must be clear during consent that open access can neither preclude any future users such as commercial organisations nor make promises that data will only be used for research purposes (Attwood and Munafò, 2016).

Those reviewing access requests and making sharing decisions should be independent, qualified and trusted (The Academy of Medical Sciences, 2013; Castell *et al.*, 2018). The composition of data access committees will depend on the resources available at each organisation holding data, and there are no set standards yet for the operation or

composition of such committees (Cheah and Piasecki, 2020), but ideally, they would contain lay or participant members (Hate *et al.*, 2015; Jao *et al.*, 2015a; Manhas *et al.*, 2015; Manhas *et al.*, 2016; Castell *et al.*, 2018). Committees also need to contain members who are qualified enough to determine whether data requestors have the required qualifications (are bona-fide researchers), whether proposed analyses are appropriate, and how difficult it would be (or how long it would take and how much it would cost) to anonymise the dataset if this has not already been done. Using a committee also means that access requests do not need to be referred back to the original researchers who may have moved on to other institutions or retired.

So that decisions around sharing are transparent (Castell *et al.*, 2018; Colombo *et al.*, 2019; NIHR, 2019; Cheah and Piasecki, 2020) organisations should publish how (and how frequently) they make sharing decisions, and who the committee members are (Institute of Medicine (IOM), 2015; Cheah and Piasecki, 2020).

Data repositories should be selected based on their ability to provide a controlled, standardised, and transparent access process. For example, the US based National Academies of Sciences explain how a multidisciplinary independent review panel assess requests for both Clinical Study Data Request (CSDR) and Vivli ClinicalStudyDataRequest.com., 2021 (ClinicalStudyDataRequest.com, 2020; National Academies of Sciences and Medicine, 2020; Vivli, 2021). ReShare, suitable for clinical trials data, also uses controlled (*"safeguarded"*) access (UK Data Service, 2021).

As identified briefly in the grey literature, access can be further controlled by preventing data sets from leaving the original researchers at all. After signing a data sharing agreement, ClinicalStudyDataRequest allows secondary researchers to perform analysis using standard software available on their website, but users are not permitted to download the data (ClinicalStudyDataRequest.com, 2021). Alternatively, operating outside of a single repository, DataSHIELD is open-source software that allows researchers to access and analyse multiple datasets simultaneously without that data ever having to leave the host institutions (Gaye *et al.*, 2014).

6.7.6 Recommendation 6: Guidance and research should stop suggesting re-consent as an option

One-off, properly informed (i.e.: with adequate information on possible future research data sharing and security measures in place) consent at the point of joining the original research study reduces the need for burdensome re-consent processes that might be beyond the capacity of already stretched researchers and intrusive for participants. To effectively re-consent or 'opt-in' the entirety of the original study population each time a sharing request is made, contact details of all participants need to be kept up to date, a task which places burden on both researchers and participants. Some participants will be uncontactable or may have died since consenting to take part in the primary study.

By requiring re-consent of participants before their data are shared, researchers will be left with a biased sample (those who were contactable, and those who agreed to the proposed secondary use) that differs to the primary dataset, limiting reproducibility and meta-analyses (Lowrance, 2002; Institute of Medicine (IOM), 2015; Medical Research Council, 2017). Excluding participants who decline after re-consent from datasets requires additional time on part of the researcher to prepare the dataset for secondary usage, and risks missing stringent targets for timely sharing, such as those proposed by the International Committee of Medical Journal Editors (ICMJE) (Taichman *et al.*, 2016).

Unfortunately, there may always be participants who decline to give consent for all future sharing at the original consent (Kass *et al.*, 2003), and it is arguably easier to exclude these participants who decline to share from the outset than it is excluding them after each re-consent. However, this approach would still result in bias in the sample of data shared for secondary use which would differ from that analysed for the primary research study.

6.7.7 Recommendation 7: Researchers should treat data as if it belongs to participants

Research data are usually owned by funders, with a whole host of other potential candidates and stakeholders (researchers, host institutions, sponsors or journals) (Cleary *et al.*, 2013). Nonetheless, if when making sharing decisions, researchers can imagine that the participants own the data, they may respect it, manage it, and use it in the way that participants would want. In other words, researchers should consider: who is the moral owner of the data? Therefore, researchers should take measures (see Recommendation 3 above) to avoid sharing the data for research that is outside of the scope of the original consent and share it

only for projects that align with participants' values (Colombo *et al.*, 2019; Mozersky *et al.*, 2020). Projects may be considered not aligned with participants' values if they are vastly different to those given as examples during the consent process and subsequently consented to by participants. It is also imperative that researchers should ensure that data are adequately anonymised and stored securely to protect privacy.

6.7.8 Recommendation 8: Provide feedback on when and to what end data have been shared

Briefly, in the spirit of transparency, researchers need to commit to providing feedback on use of participants data for secondary research. Engaging participants by keeping them informed of the types of studies using their data, and the resultant outcomes enforces researcher accountability, but is also said to increase public transparency and trust (Jao *et al.*, 2015a; Aitken *et al*., 2016b). Researchers should let participants know how their data has been used, for ethical reasons, but also to provide participants with success stories or *"positive messages about how data is used"* (Aitken *et al*., 2016b, p. 178). Participants want to know that research is benefitting them (Damschroder *et al*., 2007). Researchers may choose to use email, post a periodic newsletter or update a study or organisational website (which may be more suitable once the original study ends), but they should commit at the consent stage to the medium that the feedback will take. Informing participants at the outset of the communications they can expect to receive prevents participants feeling burdened by further contact and also gives them the opportunity to opt out of communications.

6.8 Areas for future research

Although the questionnaire was useful in determining the majority opinion, it could not explain the reasons behind these opinions, which could be further explored with qualitative methods. Some of the data gathered in the questionnaire needs verifying with further research, for example, the systematic review did not identify any data which explored whether or not knowing about sharing would discourage participants from taking part in research.

Even though recommendations for researchers have been made here based upon the available evidence gathered from the systematic review, grey literature review and finally the questionnaire, it would be prudent to further test these recommendations prior to

implementation, either with further qualitative research studies focussed specifically on the content of the recommendations or with co-production work. Organisations may wish to consider testing the relevance and acceptance of their own data sharing polices with the participants they concern. Few of the guidance documents included in the grey literature review provided recommendations on how to allay participant concerns around data sharing, and therefore future guidance should include steps specifically designed, not just to meet funder requirements but to respect participants' preferences and provide reassurance. Some areas that require further research or discussion, according to the data gathered in this study are outlined briefly below.

6.8.1 Statements for inclusion in consent forms

As recommended above (Recommendation 1) researchers should already be including in consent forms and patient information sheets brief descriptions of what data sharing is, who data may be shared with (see Recommendation 3) and how this is of benefit to research (and subsequently participants). Included in the description of what sharing is, should be an explanation of how data will be anonymised prior to sharing. These statements will need to be brief but impactful and could be developed with participants using co-production techniques. Including participants in development of study materials also moves them away from answering hypotheticals about sharing scenarios towards becoming actual stakeholders in the process (Shah *et al.*, 2018). Their inclusion ensures that statements developed and subsequently utilised are understandable for participants, but also that they place emphasis on the aspects of sharing that will reassure participants who may have concerns about data sharing.

6.8.2 Ownership

Although ownership appeared briefly in the grey literature and in the systematic review, and questionnaire respondents concluded that they thought the patients themselves should own the data, it was clear that there was not as much information available on ownership as on other aspects of sharing. It has been previously cited that researchers themselves are often unsure who owns 'their' data (Hrynaszkiewicz and Altman, 2009). This aspect of the questionnaire survey feels un-resolved and it would be useful to include more questions in future surveys regarding perceived ownership of data and who has responsibility for it. This could then be included in consent forms as a brief explanatory statement for participants.

6.8.3 Storage and access models

Even after the systematic review and questionnaire survey, there is still less known about participants' views of data storage and access than about aspects of sharing such as consent and preferred recipients of data. The scoping focus group and cognitive interviewees found questions about storage and access types less engaging, and perhaps uninteresting.

Further research should be conducted regarding participant preferences for repository types and access models. This area remains difficult even for researchers to navigate, as there are no hard rules or restrictions (Taichman *et al.*, 2016) regarding the location of shared data in a *"heterogeneous research data repository landscape"* (Pampel *et al.*, 2013, p. 1); the data can be placed in a variety of locations ranging from supplementary material in a journal to within an institution's repository or in a recognised (disciplinary) archive (Whyte, 2015).

What would be interesting, and requires further attention, is whether participants see any difference between depositing data with a journal, in a purpose-built repository, or keeping it with the original researcher and regarding which of these organisations would make access decisions. The results of research on these issues may help researchers with their future storage choices or inform researchers how best to explain repository storage to participants on consent forms. Funders and journal editors should also take account of participants' preferences when setting out their recommendations or requirements to researchers and uphold their commitment to making 'exceptions' (Taichman *et al.*, 2016) for participants' sakes when necessary.

6.8.4 Explaining and seeking consent for secondary uses of data

Researchers are likely to need guidance on how to construct some succinct and clear statements for inclusion in PIS or consent forms regarding likely secondary uses of data as well as an assessment of the type of secondary research that would be acceptable for their specific patient populations. Having participant-approved phrases, vignettes and descriptions of secondary research types may also reduce withholding of consent by participants who were concerned about sharing. This could be achieved with co-production techniques, where participants themselves contribute to design, improving the experience of those taking part, as recommended by funders and researchers (Crawford *et al.*, 2002; National Academies of Sciences and Medicine, 2020; HRA, 2021; NIHR, 2021).

More data are also required on whether there are specific concerns and challenges regarding data sharing with students, as this was a group who received fewer positive responses in the questionnaire survey and were only mentioned in two of the papers included in the systematic review (Mozersky *et al.*, 2020; Hate *et al.*, 2015). This is important, as students may be ideal candidates for conducting secondary analyses of existing data, but there is no known literature exploring this idea.

It would also be useful for researchers to ask participants their sharing and consent preferences where no explicit consent for sharing exists, for example in older studies set up before obtaining consent to share was commonplace.

6.8.5 Measuring use and misuse of shared data

Some participants in the systematic review suggested that researchers or access committees be held accountable for sharing decisions (Manhas *et al*., 2015) with penalties for data breaches or misuse (Manhas *et al*., 2015; Colombo et al., 2019). Participants outside of research included in the systematic review for this study have also suggested researchers be responsible and accountable for use of research data (Damschroder *et al*., 2007). The National Academies refer to 'metrics' on use of data for secondary research which can collect information on how data are being used including requests, approvals and publications and therefore measure the *"benefits and values"* of sharing (National Academies of Sciences and Medicine, 2020, p. 5). Analysis of such metrics could provide insights into the extent to which data are being shared and further analysed, to indicate whether the assumed benefits are being realised.

Tudur-Smith *et al* also suggest that data requests, their outcomes and any reasons for refusing to share should be made publicly available (Tudur-Smith *et al.*, 2015). Exploring reasons for rejection of requests to share, or any identified misuses of data, could also provide insight into the risks of the process.

Finally, more research is needed to identify whether penalties for misuse is something suggested off-the-cuff by just a few participants, or something that requires more widespread implementation, and how accountability might be achieved.

6.9 Questions remaining

6.9.1 Clarity on anonymisation, pseudonymisation and GDPR

All evidence identified the importance of privacy, security and subsequently, anonymisation for participants. In the patient information sheet (PIS) for the questionnaire survey I attempted to explain anonymisation and questions asked respondents for their views specifically of sharing anonymised data where personal identifiers had been removed. The papers in the systematic review also referred primarily to sharing of anonymised data, although some participants still made reference to identifiable data (Mozersky *et al.*, 2020) and some needed clarification or reassurance that their data would be anonymised before sharing (Mursaleen *et al.*, 2017a; Mursaleen *et al.*, 2017b; Cheah *et al.*, 2018; Shah *et al.*, 2018).

Most of the grey literature refers to sharing of anonymised data, although some guidance documents did not explicitly mention anonymisation at all (HEFCE *et al.*, 2016; Medical Research Council, 2016; Cancer Research UK, 2017; NIHR, 2019) and others referred more obliquely to protecting confidentiality or identity (UK Research and Innovation, 2018). In what can seem confusing at first, the ESRC guidance (ESRC, 2015) asks researchers to get consent for sharing or to anonymise data and share it (as anonymised data can be shared legally without consent as covered by the ICO (Information Commissioners Office, 2019a)). Kalkman *et al* (Kalkman *et al.*, 2019b; Kalkman *et al.*, 2019c) have also noted the use of interchangeable terms such as 'anonymised', 'anonymous' and 'de-identified' can cause confusion for researchers and presumably participants. We should also add to this, 'pseudonymised' and 'personal data' or 'identifiers'.

More than ever before, participants and members of the public are aware of their rights regarding data use and sharing (Shah *et al.*, 2018; Strycharz *et al.*, 2020) under the (relatively) new GDPR regulation (Information Commissioner's Office, 2018), (Shah *et al.*, 2018) in a new *"era of individuals' empowerment and shared-decision making"* (Karampela *et al.*, 2019, p. 6509). But as GDPR does not apply to anonymised data, which researchers could legally share in the public interest without consent or consequence, it is hard to see how GDPR applies to most researchers planning to share anonymised datasets (Kalkman *et al.*, 2019b), although, morally, consent for sharing should be sought. GDPR does apply to pseudonymised original datasets containing potentially identifying data such as date of birth. To ensure transparency, this distinction needs somehow to be explained to participants. It

would also be helpful for researchers if guidance specified whether the directions given were for anonymised data, pseudonymised data or both. Given the importance of data security to participants, there should not be any ambiguity in future guidance about anonymisation of data prior to sharing.

6.9.2 Conflict between data retention and sharing

One point that arose briefly in the grey literature was for how long data should or would be stored for sharing (Lowrance, 2002; ESRC, 2015). In accordance with the FAIR principles (Wilkinson *et al.*, 2016) data must be as accessible as possible, and for example, Research Councils UK expect that data should be accessible for *"at least ten years after publication"* (UK Research and Innovation, 2018, p. 4).

What was not identified in the literature is that this is in contrast to the remainder of the study documentation, such as consent forms, questionnaires, and crucially the link between the pseudonymised original dataset and the anonymised data. This documentation will only be held as long as is recommended by the study sponsor or the relevant ethics committee, usually only as long as necessary and with timescales communicated to participants. For example, ten years is recommended for MRC funded studies (MRC, 2017). Although a decade may seem a long time, this must be compared to the indefinite amount of time data could sit in a repository, therefore, the shared anonymised dataset will have a longer lifespan than the original dataset.

This could cause an issue should a participant wish to withdraw from sharing. Once the link between the study ID of the pseudonymised data set and the study ID of the anonymised data set has been destroyed (with or before the original data), withdrawal would not be possible. If the new EU clinical trials Regulation is adopted the UK (EMA, 2021), and researchers are required to store all original study data for an archive period of 25 years, this disparity between shared data and original study data may lessen for clinical trials at least. But for now, participants should be informed about the point at which they can no longer withdraw from sharing, or perhaps, they should be told that it is never possible to withdraw once their data has been shared?

Future researchers should find a comprehensible way to let participants know how long their original data set will be held, in what format and location, the point at which they can no longer withdraw consent for sharing, and for how long their anonymised data will be shared

(presumably indefinitely). Future *research* should ask participants their views on storage and sharing time.

6.10 Concluding remarks:

In the current global drive to accommodate data sharing from the outset of studies (Walport and Brest, 2011; PLOS., 2014; Institute of Medicine (IOM), 2015; Loder and Groves, 2015; Taichman *et al*., 2016) this study provides evidence of research participants' concerns and preferences, which, if acknowledged by researchers and funders, will ensure that advances in data sharing align with the values of the participants who contribute data. Participants need to have as much trust in the secondary research conducted as they do in the original research, and researcher's, funders and policy makers should pay particular attention to an informed consent and controlled data access as identified in the recommendations made here.

Participants have been clear about their conditions for data sharing, and research should move away from a culture of vague consent, which does not permit assessment by participants of how their data will be re-used, towards one of transparency and working with participants rather than dictating to them. For example, the recommendations presented here, although based on evidence, still require further agreement with participants before being implemented.

Further research is required, particularly into participants' views on repositories and data storage, and further clarification is required from funders regarding the tensions between retention and sharing of data, and the conflict between patient's preferences and rights regarding anonymised data. The questionnaire survey appeared to identify a link between participants age and experience of taking part in research with their sharing preferences, and this should be considered when designing further research and when explaining sharing to future participants.

More information about the benefits of data sharing, alongside the desired governance, may increase participants willingness to share, so increasing the availability of data for secondary use.

References:

Ahram, M., Othman, A., Shahrouri, M. and Mustafa, E. (2014) 'Factors influencing public participation in biobanking', *Eur J Hum Genet*, 22(4), pp. 445-51.

Aitken, M., de St. Jorre, J., Pagliari, C., Jepson, R. and Cunningham-Burley, S. (2016a) '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), pp. 73-97.

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

Aitken, M., McAteer, G., Davidson, S., Frostick, C. and Cunningham-Burley, S., 2018. 'Public preferences regarding data linkage for Health Research: a discrete choice experiment'. *International Journal of Population Data Science*, 3(1).

Alter, G. and Vardigan, M. (2015) 'Addressing Global Data Sharing Challenges', *Journal of Empirical Research on Human Research Ethics*, 10(3), pp. 317-323.

Arksey, H. and O'Malley, L. (2005) 'Scoping studies: towards a methodological framework', *International Journal of Social Research Methodology*, 8(1), pp. 19-32.

Asai, A., Ohnishi, M., Nishigaki, E., Sekimoto, M., Fukuhara, S. and Fukui, T. (2002) 'Attitudes of the Japanese public and doctors towards use of archived information and samples without informed consent: Preliminary findings based on focus group interviews', *BMC Medical Ethics*, 3(1), pp. 1-10.

Attwood, A.S. and Munafò, M.R. (2016) 'Navigating an open road', *Journal of Clinical Epidemiology*, 70, pp. 264-266.

Audrey, S., Brown, L., Campbell, R., Boyd, A. and Macleod, J. (2016) 'Young people's views about consenting to data linkage: findings from the PEARL qualitative study', *BMC Medical Research Methodology*, 16(1), p. 34.

Bates, S.R., Faulkner, W., Parry, S. and Cunningham-Burley, S. (2010) ''How do we know it's not been done yet?!' Trust, trust building and regulation in stem cell research', *Science and Public Policy*, 37(9), pp. 703-718.

Batty, G.D., Morton, S.M.B., Campbell, D., Clark, H., Smith, G.D., Hall, M., Macintyre, S. and Leon, D.A. (2004) 'The Aberdeen Children of the 1950s cohort study: background, methods and follow-up information on a new resource for the study of life course and intergenerational influences on health', *Paediatric and Perinatal Epidemiology*, 18(3), pp. 221-239.

Berg, N. (2005) 'Non-Response Bias', in Kempf-Leonard, K. (ed.) *ENCYCLOPEDIA OF SOCIAL MEASUREMENT*. Academic Press. Available at: https://ssrn.com/abstract=1691967.

BestBETs (2012) *BETs CA Worksheets*. Available at: http://bestbets.org/home/bets-introduction.php (Accessed: 25/10/2016).

Biemer, P., P, and Lyberg, L., E, (2003) *Introduction to Survey Quality*. New Jersey: Wiley.

Bobashev, G. (2011) 'Type I Error', in Lavrakas, P.J. (ed.) *Encyclopedia of Survey Research Methods,* Thousand Oaks, California: Sage Publications.

Borgman, C.L. (2012) 'The conundrum of sharing research data', *Journal of the American Society for Information Science and Technology*, 63(6), pp. 1059-1078.

Boulton, G., Rawlins, M., Vallance, P. and Walport, M. (2011) 'Science as a public enterprise: the case for open data', *The Lancet*, 377(9778), pp. 1633-1635.

Bouter, L.M. (2016) 'Open data are not enough to realize full transparency', *Journal of Clinical Epidemiology*, 70, pp. 256-257.

Boyatzis, R.E. (1998) *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA, US: Sage Publications, Inc.

Boyd, A., Golding, J., Macleod, J., Lawlor, D.A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S. and Davey Smith, G. (2013) 'Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children', *Int J Epidemiol*, 42(1), pp. 111-27.

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77-101.

Braun, K.L., Tsark, J.U., Powers, A., Croom, K., Kim, R., Gachupin, F.C. and Morris, P. (2014) 'Cancer patient perceptions about biobanking and preferred timing of consent', Biopreserv Biobank, 12(2), pp. 106-12.

Bray, I., Noble, S., Robinson, R., Molloy, L. and Tilling, K. (2017) 'Mode of delivery affected questionnaire response rates in a birth cohort study', *J Clin Epidemiol*, 81, pp. 64-71.

Broekstra, R., Aris-Meijer, J., Maeckelberghe, E., Stolk, R., Otten, S., (2020), 'Trust in Centralized Large-Scale Data Repository: A Qualitative Analysis', *Journal of Empirical Research on Human Research Ethics*, 15(4), pp. 365-378

Brown, D., Allik, M., Dundas, R., Leyland and H., A. (2014) *Carstairs Scores for Scottish Postcode Sectors, Datazones and Output Areas from the 2011 Census*. University of Glasgow: MRC/CSO Social and Public Health Sciences Unit.

Bujang, M.A., Sa'at, N., Sidik, T.M.I.T.A.B. and Joo, L.C. (2018) 'Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data', *The Malaysian journal of medical sciences: MJMS*, 25(4), pp. 122-130.

Bull, S., Roberts, N. and Parker, M. (2015) 'Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public Health Research: A Systematic Scoping Review', *Journal of Empirical Research on Human Research Ethics*, 10(3), pp. 225-38.

Campbell, B., Thomson, H., Slater, J., Coward, C., Wyatt, K. and Sweeney, K. (2007) 'Extracting information from hospital records: what patients think about consent', Qual Saf Health Care, 16(6), pp. 404-8.

Cancer Research UK (2017) *Cancer Research UK Policy on Data Sharing and Preservation*. [Online]. Available at: https://www.cancerresearchuk.org/sites/default/files/cruk_data_sharing_policy_2020_final.pdf (Accessed: 08/12/2021).

Carr, D. and Littler, K. (2015) 'Sharing Research Data to Improve Public Health: A Funder Perspective', *Journal of Empirical Research on Human Research Ethics*, 10(3), pp. 314-316.

Caruana, E.J., Roman, M., Hernández-Sánchez, J. and Solli, P. (2015) 'Longitudinal studies', *Journal of Thoracic Disease*, 7(11), pp. E537-E540.

Casey, J.A., Schwartz, B.S., Stewart, W.F. and Adler, N.E. (2016) 'Using Electronic Health Records for Population Health Research: A Review of Methods and Applications', *Annual Review of Public Health*, 37(1), pp. 61-81.

CASP (2013) *Critical Appraisal Skills Framework Qualitative Appraisal Tool*. Available at: http://www.casp-uk.net/#!casp-tools-checklists/c18f8 (Accessed: 08/12/2021).

Castell, S., Bukowski, G., Burkitt, R. and Rossington, T. (2018) *Consent to use human tissue and linked health data in health research,* Human Research Authority and Human Tissue Authority, [Online] Available at: https://s3.eu-west-2.amazonaws.com/www.hra.nhs.uk/media/documents/Consent_to_use_human_tissue_and_linked_health_data_in_health_research_FINAL.pdf (Accessed 08/12/2021).

Centre for Reviews and Dissemination (CRD) (2013) *Systematic Reviews: CRD's guidance for undertaking reviews in health care*. 3rd edn., York: Centre for Reviews and Dissemination, York University.

CHAIN (Contact, Help, Advice and Information Network (2022), Available at: https://www.chain-network.org.uk/ (Accessed 14/04/2022).

Chan, T.W., Mackey, S. and Hegney, D.G. (2012) 'Patients' experiences on donation of their residual biological samples and the impact of these experiences on the type of consent given for the future research use of the tissue: a systematic review', *Int J Evid Based Healthc*, 10(1), pp. 9-26.

Chawinga, W.D. and Zinn, S. (2019) 'Global perspectives of research data sharing: A systematic literature review', *Library & Information Science Research*, 41(2), pp. 109-122.

Cheah, P.Y., Jatupornpimol, N., Hanboonkunupakarn, B., Khirikoekkong, N., Jittamala, P., Pukrittayakamee, S., Day, N.P.J., Parker, M. and Bull, S. (2018) 'Challenges arising when seeking broad consent for health research data sharing: a qualitative study of perspectives in Thailand', *BMC Medical Ethics*, 19(1), p. 86.

Cheah, P.Y. and Piasecki, J. (2020) 'Data Access Committees', *BMC Medical Ethics*, 21(1), pp. 12-20.

Cheah, P.Y., Tangseefa, D., Somsaman, A., Chunsuttiwat, T., Nosten, F., Day, N.P., Bull, S. and Parker, M. (2015) 'Perceived Benefits, Harms, and Views About How to Share Data Responsibly: A Qualitative Study of Experiences With and Attitudes Toward Data Sharing Among Research Staff and Community Representatives in Thailand', *J Empir Res Hum Res Ethics*, 10(3), pp. 278-89.

Cleary, M., Jackson, D. and Walter, G. (2013) 'Research data ownership and dissemination: is it too simple to suggest that 'possession is nine-tenths of the law'?', *Journal of Clinical Nursing*, 22(15-16), pp. 2087-2089.

Clerkin, P., Buckley, B.S., Murphy, A.W. and MacFarlane, A.E. (2013) 'Patients' views about the use of their personal information from general practice medical records in health research: a qualitative study in Ireland', *Fam Pract*, 30(1), pp. 105-12.

ClinicalStudyDataRequest.com (2020) *Clinical Study Data Request*. Available at: https://www.clinicalstudydatarequest.com/Default.aspx (Accessed: 08/12/2021).

ClinicalStudyDataRequest.com (2021) *How it works-access to data*. Available at: https://www.clinicalstudydatarequest.com/Help/Help-How-to-Request-Data.aspx (Accessed: 08/12/2021).

Cochrane, A., Welch, C., Fairhurst, C., Cockayne, S., Torgerson, D.J. and Group, O.S. (2020) 'An evaluation of a personalised text message reminder compared to a standard text message on postal questionnaire response rates: an embedded randomised controlled trial', *F1000Research*, 9, pp. 154-154.

COGGON, D., ROSE, G. A., & BARKER, D. J. P. (2003) 'Longitudinal Studies', in Epidemiology for the uninitiated. BMJ books. Available at: https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated.

Colombo, C., Roberto, A., Krleza-Jeric, K., Parmelli, E. and Banzi, R. (2019) 'Sharing individual participant data from clinical studies: a cross-sectional online survey among Italian patient and citizen groups', *BMJ Open*, 9(2), pp. bmjopen-2018-024863.

Cooper, I.D. and Johnson, T.P. (2016) 'How to use survey results', *Journal of the Medical Library Association,* 104(2), pp. 174-177.

Corti, L., Van den Eynden, V., Bishop, L. and Woollard, M. (2014) *Managing and Sharing Research Data*. London: Sage.

Courbier, S., Dimond, R. and Bros-Facer, V. (2019) 'Share and protect our health data: an evidence-based approach to rare disease patients' perspectives on data sharing and data protection - quantitative survey and recommendations', *Orphanet Journal of Rare Diseases*, 14(1), p. 175.

Crawford, M.J., Rutter, D., Manley, C., Weaver, T., Bhui, K., Fulop, N. and Tyrer, P. (2002) 'Systematic review of involving patients in the planning and development of health care', *BMJ*, 325(7375), p. 1263.

Creswell, J.W. and Plano Clark, V.L. (2018) *Designing and conducting mixed methods research,* California: Sage.

Cunningham, M. and Wells, M. (2017) 'Qualitative analysis of 6961 free-text comments from the first National Cancer Patient Experience Survey in Scotland', *BMJ Open*, 7(6), p. e015726.

Damschroder, L.J., Pritts, J.L., Neblo, M.A., Kalarickal, R.J., Creswell, J.W. and Hayward, R.A. (2007) 'Patients, privacy and trust: patients' willingness to allow researchers to access their medical records', *Social Science and Medicine*, 64(1), pp. 223-35.

da Silva, M.E., Coeli, C.M., Ventura, M., Palacios, M., Magnanini, M.M., Camargo, T.M. and Camargo, K.R., Jr. (2012) 'Informed consent for record linkage: a systematic review', *J Med Ethics*, 38(10), pp. 639-42.

Data Protection Commissioner, E.P. (1995) *95/46/EC - The Data Protection Directive* Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (Accessed: 08/12/2021).

DCC (2015) *A world leading centre of expertise in digital information curation*. Available at: www.dcc.ac.uk (Accessed: 08/12/2021).

de Leeuw, E.D. (2001) 'Reducing Missing Data in Surveys: An Overview of Methods', *Quality and Quantity*, 35(2), pp. 147-160.

De Vaus, D. (2014) *Surveys in social research*. 6th edn. Oxon: Routledge.

Deeks, J., Higgins, JPT., Altman, DG., (2021) 'Chapter 10: Analysing data and undertaking meta-analyses', in Higgins JPT, T.J., Chandler J, Cumpston M, Li T, Page MJ, Welch VA (ed.) *Cochrane Handbook for Systematic Reviews of Interventions version 6.2*., Available at: https://training.cochrane.org/handbook/current/chapter-10 (Accessed 08/12/2021).

Devereaux, P.J. (2019) 'Access to clinical trial data—Commentary', *Clinical Trials*, 16(5), pp. 552-554.

Dhand, N.K., & Khatkar, M. S. (2014) *Statulator: An online statistical calculator. Sample Size Calculator for Estimating a Single Proportion.* Available at: http://statulator.com/SampleSize/ss1P.html (Accessed: 08/12/2021).

Eaden, J., Mayberry, M.K. and Mayberry, J.F. (1999) 'Questionnaires: the use and abuse of social survey methods in medical research', *Postgrad Med J*, 75(885), pp. 397-400.

Editorial (2018) 'Data sharing and the future of science', *Nature Communications*, 9(1), pp. 2817-2818.

Edmonds, W. and Kennedy, T. (2017) 'An Applied Guide to Research Designs: Quantitative, Qualitative, and Mixed Methods', in SAGE Publications, Inc. Available at: https://methods.sagepub.com/book/an-applied-guide-to-research-designs-2e (Accessed: 08/12/2021).

Edwards, P.J., Roberts, I., Clarke, M.J., DiGuiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L.M. and Pratap, S. (2009) 'Methods to increase response to postal and electronic questionnaires', *Cochrane Database of Systematic Reviews*, 18(2), Available at: https://pubmed.ncbi.nlm.nih.gov/17443629/ (Accessed 08/12/2021).

EMA (2021) *Clinical Trials regulation*. Available at: https://www.ema.europa.eu/en/human-regulatory/research-development/clinical-trials/clinical-trials-regulation (Accessed: 08/12/2021).

EndNote (2021) *EndNote 20*. Available at: https://endnote.com/product-details (Accessed: 08/12/2021).

ESRC (2015) *ESRC expectations on Research Data Management and Sharing- ESRC Research Data Policy*. Available at: https://www.ukri.org/wp-content/uploads/2021/07/ESRC-200721-ResearchDataPolicy.pdf (Accessed: 08/12/2021).

Fereday, J. and Muir-Cochrane, E. (2006) 'Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development', *International Journal of Qualitative Methods*, 5(1), pp. 80-92.

Fink, A. (2003a) *How to Sample in Surveys*. Available at: https://methods.sagepub.com/book/how-to-sample-in-surveys (Accessed: 08/12/2021).

Fink, A. (2003b) *The survey kit*. 2nd edn. California: Sage Publications.

Flowerdew, R., & Martin, D.M. (Eds.) (2005) *Methods in Human Geography: A guide for students doing a research project*. Routledge.

Fuse (2022) *Fuse: The Centre for Translational Research in Public Health*. Available at: http://www.fuse.ac.uk/ (Accessed: 20/05/2022).

Garrison, N.A., Sathe, N.A., Antommaria, A.H., Holm, I.A., Sanderson, S.C., Smith, M.E., McPheeters, M.L. and Clayton, E.W. (2016) 'A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States', *Genet Med*, 18(7), pp. 663-71.

Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E.M., Minion, J., Boyd, A.W., Newby, C.J., Nuotio, M.-L., Wilson, R., Butters, O., Murtagh, B., Demir, I., Doiron, D., Giepmans, L., Wallace, S.E., Budin-Ljøsne, I., Oliver Schmidt, C., Boffetta, P., Boniol, M., Bota, M., Carter, K.W., deKlerk, N., Dibben, C., Francis, R.W., Hiekkalinna, T., Hveem, K., Kvaløy, K., Millar, S., Perry, I.J., Peters, A., Phillips, C.M., Popham, F., Raab, G., Reischl, E., Sheehan, N., Waldenberger, M., Perola, M., van den Heuvel, E., Macleod, J., Knoppers, B.M., Stolk, R.P., Fortier, I., Harris, J.R., Woffenbuttel, B.H., Murtagh, M.J., Ferretti, V. and Burton, P.R. (2014) 'DataSHIELD: taking the analysis to the data, not the data to the analysis', *International Journal of Epidemiology*, 43(6), pp. 1929-1944.

Gelman, A. and Loken, E. (2014) 'The Statistical Crisis in Science', *American Scientist*, 102(6), p.460.

Gibbs, A. (1997) 'Focus Groups', *Social Research Update, Sociology at Surrey*, Winter 1997(19).

Godlee, F. and Groves, T. (2012) '*The new BMJ policy on sharing data from drug and device trials', BMJ, 20(345)*.

Graves, A., McLaughlin, D., Leung, J. and Powers, J. (2019) 'Consent to data linkage in a large online epidemiological survey of 18–23 year old Australian women in 2012–13', *BMC Medical Research Methodology*, 19(1), pp. 235-244.

Green & Thorogood (2014) *Qualitative Methods for Health Research,* 3rd edition., London, Sage.

Grix, J. (2010) *The Foundations of Research*. 2nd edition., London, Palgrave Macmillan.

Guest, G., MacQueen, K., M., and Namey, E., E., (2012) *Applied Thematic Analysis*. Available at: https://methods.sagepub.com/book/applied-thematic-analysis (Accessed: 08/12/2021).

Haigh, F., Kemp, L., Bazeley, P. and Haigh, N. (2019) 'Developing a critical realist informed framework to explain how the human rights and social determinants of health relationship works', *BMC Public Health*, 19(1), pp. 1571-1583.

Hajduk, G.K., Jamieson, N.E., Baker, B.L., Olesen, O.F. and Lang, T. (2019) 'It is not enough that we require data to be shared; we have to make sharing easy, feasible and accessible too!', *BMJ Global Health*, 4(4), p. e001550.

Hall, R., Frederick., (2013) 'Chapter 7: Mixed Methods: In search of a paradigm,' in Thao Le and Quynh Le (eds.) Conducting Research in a Changing and Challenging World, New York, Nova Science Publishers Inc.

Hannes, K. and Lockwood, C. (2011) *Synthesizing Qualitative Research: Choosing the Right Approach*, Wiley Blackwell, Oxford.

Hate, K., Meherally, S., Shah More, N., Jayaraman, A., Bull, S., Parker, M. and Osrin, D. (2015) 'Sweat, Skepticism, and Uncharted Territory: A Qualitative Study of Opinions on Data Sharing Among Public Health Researchers and Research Participants in Mumbai, India', *Journal of Empirical Research on Human Research Ethics*, 10(3), pp. 239-250.

Haug, C.J. (2017) 'Whose Data Are They Anyway? Can a Patient Perspective Advance the Data-Sharing Debate?', *New England Journal of Medicine*, 376(23), pp. 2203-2205.

Heale, R. and Forbes, D. (2013) 'Understanding triangulation in research', *Evidence Based Nursing*, 16(4), pp. 98-98.

Health Research Authority (2019) *Informing participants and seeking consent*. Available at: http://www.hra.nhs.uk/resources/before-you-apply/consent-and-participation/consent-and-participant-information/ (Accessed: 08/12/2021).

HEFCE, RCUK, Universities UK and & Wellcome Trust (2016) *Concordat on Open Research Data*. [Online]. Available at: https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/ (Accessed: 08/12/2021).

Hendrix, K.S., Meslin, E.M., Carroll, A.E. and Downs, S.M. (2013) 'Attitudes about the use of newborn dried blood spots for research: a survey of underrepresented parents', *Acad Pediatr*, 13(5), pp. 451-7.

Hidalgo-Landa, A., Szabo, I., Le Brun, L., Owen, I., and Fletcher, G. (2011) 'Evidence based Scoping Reviews', *The Electronic Journal Information Systems Evaluation*, 14(1), pp. 46-52.

Higgins JPT, T.J., Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) (2021) *Cochrane Handbook for Systematic Reviews of Interventions 2nd edition*. Cochrane, 2021. Available at: www.training.cochrane.org/handbook.

Hill, E., Turner, E., Martin, R. and Donovan, J. (2013) 'Let's get the best quality research we can': public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study', *BMC Medical Research Methodology*, 13(1), pp. 72-82.

Hoeyer, K., Olofsson, B.O., Mjörndal, T. and Lynöe, N. (2004) 'Informed consent and biobanks: a population-based study of attitudes towards tissue donation for genetic research', *Scand J Public Health*, 32(3), pp. 224-9.

Hopkins, C., Sydes, M., Murray, G., Woolfall, K., Clarke, M., Williamson, P. and Tudur-Smith, C. (2016) 'UK publicly funded Clinical Trials Units supported a controlled access approach to share individual participant data but highlighted concerns', *Journal of Clinical Epidemiology*, 70, pp. 17-25.

Howe, N., Giles, E., Newbury-Birch, D. and McColl, E. (2018) 'Systematic review of participants' attitudes towards data sharing: a thematic synthesis', *Journal of Health Services Research & Policy*, 23(2), pp. 123-133.

HRA (2017) *Applying a proportionate approach to the process of seeking consent* (V1.01). Available at: https://s3.eu-west-2.amazonaws.com/www.hra.nhs.uk/media/documents/Proportionate_approach_to_seeking_consent_HRA_Guidance.pdf (Accessed: 08/12/2021).

HRA (2018) *Data protection and information governance*. Available at: https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/data-protection-and-information-governance/ (Accessed: 08/12/2021).

HRA (2021) *Public involvement*. Available at: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/ (Accessed: 08/12/2021).

Hrynaszkiewicz, I. and Altman, D.G. (2009) 'Towards agreement on best practice for publishing raw clinical trial data', *Trials*, 10(1), pp. 17-21.

Hunter, I.M., Whiddett, R.J., Norris, A.C., McDonald, B.W. and Waldon, J.A. (2009) 'New Zealanders' attitudes towards access to their electronic health records: preliminary results from a national study using vignettes', *Health Informatics J*, 15(3), pp. 212-28.

Hutchings, E., Loomes, M., Butow, P. and Boyle, F.M. (2020) 'A systematic literature review of health consumer attitudes towards secondary use and sharing of health administrative and clinical trial data: a focus on privacy, trust, and transparency', *Systematic Reviews*, 9(1), pp. 235-276.

Information Commissioner's Office (2018) *The Data Protection Act 2018*. Available at: http://www.legislation.gov.uk/ukpga/2018/12/contents/enacted (Accessed: 08/12/2021).

Information Commissioners Office (2011) *Data Sharing Code of Practice,* Available at: https://ico.org.uk/media/for-organisations/documents/1068/data_sharing_code_of_practice.pdf (Accessed: 08/12/2021).

Information Commissioners Office (2019a) *Data Sharing Code of Practice- draft code for consultation,* Available at: https://ico.org.uk/media/about-the-ico/consultations/2615361/data-sharing-code-for-public-consultation.pdf (Accessed: 08/12/2021).

Information Commissioners Office (2019b) *What is personal data?* Available at: http://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/ (Accessed: 08/12/2021).

Information Commissioners Office (2020) *Data Sharing Code of Practice*. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/data-sharing-a-code-of-practice/ (Accessed: 08/12/2021).

Innes, N.P.T., Clarkson, J.E., Speed, C., Douglas, G.V.A., Maguire, A. and Fi, C.T.C. (2013) 'The FiCTION dental trial protocol – filling children's teeth: indicated or not?', *BMC Oral Health*, 13(1), pp. 25-38.

Institute of Medicine (IOM) (2015) 'Sharing Clinical Trial Data maximising benefits minimising risk', JAMA, 13(8), pp. 793-794.

Institute of Medicine (US) (2013) *Sharing Clinical Research Data: Workshop Summary.* National Academies Press (US). Available at: https://books.google.co.uk/books?hl=en&lr=&id=fhF1AgAAQBAJ&oi=fnd&pg=PR21&dq=Sha

ring+Clinical+Research+Data:+Workshop+Summary.&ots=EiBIgQSfUD&sig=EVssbHtytGmd791q-P2f4fSaevE#v=onepage&q=Sharing%20Clinical%20Research%20Data%3A%20Workshop%20Summary.&f=false (Accessed 08/12/2021).

Ivankova, N.V., Creswell, J.W. and Stick, S.L. (2006) 'Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice', *Field Methods*, 18(1), pp. 3-20.

Jao, I., Kombe, F., Mwalukore, S., Bull, S., Parker, M., Kamuya, D., Molyneux, S. and Marsh, V. (2015a) 'Involving Research Stakeholders in Developing Policy on Sharing Public Health Research Data in Kenya: Views on Fair Process for Informed Consent, Access Oversight, and Community Engagement', *J Empir Res Hum Res Ethics*, 10(3), pp. 264-77.

Jao, I., Kombe, F., Mwalukore, S., Bull, S., Parker, M., Kamuya, D., Molyneux, S. and Marsh, V. (2015b) 'Research Stakeholders' Views on Benefits and Challenges for Public Health Research Data Sharing in Kenya: The Importance of Trust and Social Relations', *PLoS One*, 10(9), p. e0135545.

Johnson, R.B., Onwuegbuzie, A.J. and Turner, L.A. (2007) 'Toward a Definition of Mixed Methods Research', *Journal of Mixed Methods Research*, 1(2), pp. 112-133.

Joly, Y., Dalpé, G., So, D. and Birko, S. (2015) 'Fair Shares and Sharing Fairly: A Survey of Public Views on Open Science, Informed Consent and Participatory Research in Biobanking', *PLOS ONE*, 10(7), p. e0129893.

Kalkman, S., van Delden, J., Banerjee, A., Tyl, B., Mostert, M. and van Thiel, G. (2019a) 'Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence', *Journal of Medical Ethics*, doi: 10.1136/medethics-2019-105651.

Kalkman, S., Mostert, M., Udo-Beauvisage, N., van Delden, J.J. and van Thiel, G.J. (2019b) 'Responsible data sharing in a big data-driven translational research platform: lessons learned', *BMC Medical Informatics and Decision Making*, 19(1), pp. 283-290.

Kalkman, S., Mostert, M., Gerlinger, C., van Delden, J.J.M. and van Thiel, G.J.M.W. (2019c) 'Responsible data sharing in international health research: a systematic review of principles and norms', *BMC Medical Ethics*, 20(1), pp. 21-34.

Karampela, M., Ouhbi, S. and Isomursu, M. (2019) *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 23-27 July 2019.

Kass, N.E., Natowicz, M.R., Hull, S.C., Faden, R.R., Plantinga, L., Gostin, L.O. and Slutsman, J. (2003) 'The Use of Medical Records in Research: What Do Patients Want?', *The Journal of Law, Medicine & Ethics*, 31(3), pp. 429-433.

Kaufman, D.J., Murphy-Bollinger, J., Scott, J. and Hudson, K.L. (2009) 'Public Opinion about the Importance of Privacy in Biobank Research', *American Journal of Human Genetics*, 85(5), pp. 643-654.

Keding, A., Brabyn, S., MacPherson, H., Richmond, S.J. and Torgerson, D.J. (2016) 'Text message reminders to improve questionnaire response rates', *J Clin Epidemiol*, 79, pp. 90-95.

Keerie, C., Tuck, C., Milne, G., Eldridge, S., Wright, N. and Lewis, S.C. (2018) 'Data sharing in clinical trials – practical guidance on anonymising trial datasets', *Trials*, 19(1), pp. 25-33.

Kessler, M.M. (1963) 'Bibliographic coupling between scientific papers', *American Documentation*, 14(1), pp. 10-25.

Kim, K.K., Joseph, J.G. and Ohno-Machado, L. (2015) 'Comparison of consumers' views on electronic data sharing for healthcare and research', *J Am Med Inform Assoc*, 22(4), pp. 821-30.

Kitzinger, J. (1995) 'Qualitative Research: Introducing focus groups', *BMJ*, 311(7000), pp. 299-302.

Koers, H. (2016) 'How do we make it easy and rewarding for researchers to share their data? A publisher's perspective', *Journal of Clinical Epidemiology*, 70, pp. 261-263.

Kouamé, J.B. (2010) 'Using Readability Tests to Improve the Accuracy of Evaluation Documents Intended for Low-Literate Participants', *Journal of MultiDisciplinary Evaluation*, 6(14), pp. 132-139.

Kreuger, R. and Casey, M., A., (2015) *Focus Group Interviewing Research Methods*. Available at: https://richardakrueger.com/focus-group-interviewing/ (Accessed: 08/12/2021).

Lemke, A.A., Wolf, W.A., Hebert-Beirne, J. and Smith, M.E. (2010) 'Public and biobank participant attitudes toward genetic research participation and data sharing', *Public Health Genomics*, 13(6), pp. 368-77.

Levac, D., Colquhoun, H. and O'Brien, K. (2010) 'Scoping studies: advancing the methodology', *Implementation Science*, 5(1), pp. 69-78.

Lo, B. (2015) 'Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk', *JAMA*, 313(8), pp. 793-794.

Loder, E. (2013) 'Sharing data from clinical trials: where we are and what lies ahead', *BMJ: British Medical Journal*, 347, p. f4794.

Loder, E. and Groves, T. (2015) 'The BMJ requires data sharing on request for all trials', *BMJ: British Medical Journal*, 350, p. h2373.

Lowrance, W.W. (2002) *Learning from Experience: Privacy and the Secondary Use of Data in Health Research*. London: The Nuffield Trust for research and policy studies in health services.

Ludman, E.J., Fullerton, S.M., Spangler, L., Trinidad, S.B., Fujii, M.M., Jarvik, G.P., Larson, E.B. and Burke, W. (2010) 'Glad you asked: Participants' Opinions of Re-Consent for dbGaP Data Submission', *Journal of empirical research on human research ethics: JERHRE*, 5(3), pp. 9-16.

Manhas, K.P., Dodd, S.X., Page, S., Letourneau, N., Adair, C.E., Cui, X. and Tough, S.C. (2018) 'Sharing longitudinal, non-biological birth cohort data: a cross-sectional analysis of parent consent preferences', *BMC Medical Informatics and Decision Making*, 18(1), p. 97.

Manhas, K.P., Page, S., Dodd, S.X., Letourneau, N., Ambrose, A., Cui, X. and Tough, S.C. (2015) 'Parent perspectives on privacy and governance for a pediatric repository of non-biological, research data', *J Empir Res Hum Res Ethics*, 10(1), pp. 88-99.

Manhas, K.P., Page, S., Dodd, S.X., Letourneau, N., Ambrose, A., Cui, X. and Tough, S.C. (2016) 'Parental perspectives on consent for participation in large-scale, non-biological data repositories', *Life Sciences, Society and Policy*, 12, pp. 1-14.

Mauthner, N.S. and Parry, O. (2013) 'Open Access Digital Data Sharing: Principles, Policies and Practices', *Social Epistemology*, 27(1), pp. 47-67.

Mazor, K.M., Richards, A., Gallagher, M., Arterburn, D.E., Raebel, M.A., Nowell, W.B., Curtis, J.R., Paolino, A.R. and Toh, S. (2017) 'Stakeholders' views on data sharing in multicenter studies', *Journal of Comparative Effectiveness Research*, 6(6), pp. 537-547.

McColl, E., Jacoby, A., Thomas, L., Soutter, J., Bamford, C., Steen, N., Thomas, R., Harvey, E., Garratt, A. and Bond, J. (2001) 'Design and use of questionnaires: a review of best practice

applicable to surveys of health service staff and patients', *Health Technol Assess*, 5(31), pp. 1-256.

McDonald, J.H. (2014) 'Chi-square test of independence', in *Handbook of Biological Statistics (3rd ed.)*, Sparky House Publishing, Baltimore, [Online]. Available at: http://www.biostathandbook.com/chiind.html (Accessed 09/06/2022).

McGuire, A.L., Hamilton, J.A., Lunstroth, R., McCullough, L.B. and Goldman, A. (2008) 'DNA data sharing: research participants' perspectives', *Genetics in Medicine*, 10(1), pp. 46-53.

McGuire, A.L., Oliver, J.M., Slashinski, M.J., Graves, J.L., Wang, T., Kelly, P.A., Fisher, W., Lau, C.C., Goss, J., Okcu, M., Treadwell-Deering, D., Goldman, A.M., Noebels, J.L. and Hilsenbeck, S.G. (2011) 'To share or not to share: a randomized trial of consent for data sharing in genome research', *Genet Med*, 13(11), pp. 948-55.

McLafferty, I. (2004) 'Focus group interviews as a data collecting strategy', *Journal of Advanced Nursing*, 48(2), pp. 187-194.

Medical Research Council (2011) *MRC Policy and guidance on Sharing of Research Data from Population and Patient Studies,* Medical Research Council. [Online]. Available at: https://mrc.ukri.org/research/policies-and-guidance-for-researchers/data-sharing/ (Accessed 08/12/2021).

Medical Research Council (2016) *MRC Policy on Research Data Sharing* www.mrc.ac.uk: Medical Research Council. [Online]. Available at: https://mrc.ukri.org/research/policies-and-guidance-for-researchers/data-sharing/ (Accessed: 08/12/2021).

Medical Research Council (2017) *Using information about people in health research,* (Version 1.0, August 2017). Available at: https://mrc.ukri.org/documents/pdf/using-information-about-people-in-health-research-2017/ (Accessed: 08/12/2021).

Mello, M.M., Francer, J.K., Wilenzick, M., Teden, P., Bierer, B.E. and Barnes, M. (2013) 'Preparing for Responsible Sharing of Clinical Trial Data', *New England Journal of Medicine*, 369(17), pp. 1651-1658.

Mello, M.M., Lieou, V. and Goodman, S.N. (2018) 'Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing', *New England Journal of Medicine*, 378(23), pp. 2202-2211.

Mendeley (2020) *Discover Mendeley Data*. Available at: https://data.mendeley.com/ (Accessed: 08/12/2021).

Merson, L., Phong, T.V., Nhan le, N.T., Dung, N.T., Ngan, T.T., Kinh, N.V., Parker, M. and Bull, S. (2015) 'Trust, Respect, and Reciprocity: Informing Culturally Appropriate Data-Sharing Practice in Vietnam', *J Empir Res Hum Res Ethics*, 10(3), pp. 251-63.

Meterko, M., Restuccia, J.D., Stolzmann, K., Mohr, D., Brennan, C., Glasgow, J. and Kaboli, P. (2015) 'Response Rates, Nonresponse Bias, and Data Quality: Results from a National Survey of Senior Healthcare Leaders', *Public Opinion Quarterly*, 79(1), pp. 130-144.

Meyer, M.N. (2018) 'Practical Tips for Ethical Data Sharing', *Advances in Methods and Practices in Psychological Science*, 1(1), pp. 131-144.

Moher D, S.L., Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. (2015) 'Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement.', *Syst Rev.*, 4(1), pp. 1-9.

MoreTrials (2017) *Research Waste – Focus on sharing individual patient data distracts from improving trial transparency and also makes it much harder to do randomised trials?* Available at: https://www.researchgate.net/publication/317805002_Focus_on_sharing_individual_patient_data_distracts_from_other_ways_of_improving_trial_transparency (Accessed: 08/12/2021).

Morgan D. L (1997) *Focus Groups As Qualitative Research*. Thousand Oaks, California: Sage.

Mourby, M.J., Doidge, J., Jones, K.H., Aidinlis, S., Smith, H., Bell, J., Gilbert, R., Dutey-Magni, P. and Kaye, J. (2019) 'Health Data Linkage for UK Public Interest Research: Key Obstacles and Solutions', *International Journal of Population Data Science*, 4(1), pp. 1093-1093.

Mozersky, J., Parsons, M., Walsh, H., Baldwin, K., McIntosh, T. and DuBois, J.M. (2020) 'Research Participant Views regarding Qualitative Data Sharing', *Ethics & Human Research*, 42(2), pp. 13-27.

MRC (2017) 'MRC Regulatory Support Centre: Retention framework for research data and records' [Online]. Available at: https://mrc.ukri.org/documents/pdf/retention-framework-for-research-data-and-records/#:~:text=For%20basic%20research%20%2D%20Research%20data,the%20study%20h

as%20been%20completed.&text=to%20keep%20data%20indefinitely%20as,use%20them%20further%20for%20research. (Accessed: 08/12/2021).

Mursaleen, L., Stamford, J., Schmidt, P., Dean, J., Windle, R., Jones, D. and Matthews, H. (2017a) 'Choices on selective clinical data sharing by people with Parkinson's disease', *Research and Reviews in Parkinsonism*, 7, pp. 29-32.

Mursaleen, L.R., Stamford, J.A., Jones, D.A., Windle, R. and Isaacs, T. (2017b) 'Attitudes Towards Data Collection, Ownership and Sharing Among Patients with Parkinson's Disease', *J Parkinsons Dis*, 7(3), pp. 523-531.

Nair, K., Willison, D., Holbrook, A. and Keshavjee, K. (2004) 'Patients' consent preferences regarding the use of their health information for research purposes: a qualitative study', *J Health Serv Res Policy*, 9(1), pp. 22-7.

National Academies of Sciences, E. and Medicine (2020) *Reflections on Sharing Clinical Trial Data: Challenges and a Way Forward: Proceedings of a Workshop*. Washington, DC: The National Academies Press.

National Learning and Work Institute (England and Wales) (2019) *SMOG Readability Calculator*. Available at: https://learningandwork.org.uk/resources/research-and-reports/readability-how-to-produce-clear-written-materials-for-a-range-of-readers/ (Accessed: 08/12/2021).

Newbury-Birch, D., Allan, K, (2019) *Co-creating and Co-producing Research Evidence: A Guide for Practitioners and Academics in Health, Social Care and Education Settings, 1st Edition*. London: Routledge.

NIHR (2016) *Good Clinical Practice (GCP)* (3.1). Available at: https://www.nihr.ac.uk/health-and-care-professionals/learning-and-support/good-clinical-practice.htm (Accessed: 08/12/2021).

NIHR (2019) *NIHR Position on the sharing of research data*. Available at: https://www.nihr.ac.uk/documents/nihr-position-on-the-sharing-of-research-data/12253?pr= (Accessed: 08/12/2021).

NIHR (2021) *Engage patients to help shape your clinical research*. Available at: https://www.nihr.ac.uk/explore-nihr/industry/pecd.htm (Accessed: 08/12/2021).

NIHR (2022a) *Clinical Trial Guide*. Available at: https://www.nihr.ac.uk/documents/clinical-trials-guide/20595 (Accessed: 25/05/2022).

NIHR (2022b) *Public Health Research.* Available at: https://www.nihr.ac.uk/explore-nihr/funding-programmes/public-health-research.htm (Accessed 20/05/2022).

Nomis official labour market statistics (2014) *Highest level of qualification by age,* Available at: http://www.nomisweb.co.uk/census/2011/lc5102ew (Accessed: 08/12/2021).

NVivo (2021) *NVivo Qualitative data analysis,* Available at: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home (Accessed 08/12/2021).

OECD (2007) *OECD Principles and Guidelines for Access to Research Data from Public Funding',* Available at: https://www.oecd.org/sti/inno/38500813.pdf (Accessed 08/12/2021).

Office of National Statistics (2011) *2011 Census: Key Statistics and Quick Statistics for Local Authorities in the United Kingdom*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/keystatisticsandquickstatisticsforlocalauthoritiesintheunitedkingdom/2013-10-11#population (Accessed: 08/12/2021).

Office of National Statistics (2016) *Classifications and harmonisation*. Available at: https://www.ons.gov.uk/methodology/classificationsandstandards/measuringequality/ethnicgroupnationalidentityandreligion (Accessed: 08/12/2021).

Office of National Statistics (2020) *Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2020*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2020 (Accessed: 08/12/2021).

Ohmann, C., Banzi, R., Canham, S., Battaglia, S., Matei, M., Ariyo, C., Becnel, L., Bierer, B., Bowers, S., Clivio, L., Dias, M., Druml, C., Faure, H., Fenner, M., Galvez, J., Ghersi, D., Gluud, C., Groves, T., Houston, P., Karam, G., Kalra, D., Knowles, R.L., Krleža-Jerić, K., Kubiak, C., Kuchinke, W., Kush, R., Lukkarinen, A., Marques, P.S., Newbigging, A., O'Callaghan, J., Ravaud, P., Schlünder, I., Shanahan, D., Sitter, H., Spalding, D., Tudur-Smith, C., van Reusel, P., van Veen, E.-B., Visser, G.R., Wilson, J. and Demotes-Mainard, J. (2017) 'Sharing and reuse

of individual participant data from clinical trials: principles and recommendations', *BMJ Open*, 7(12), p. e018647.

Open Research Data Task Force (2018) *Realising the potential-Final report of the Open Research Data Task Force,* Available at: https://www.gov.uk/government/publications/open-research-data-task-force-final-report, (Accessed 08/12/2021).

Oppenheim, A.N. (1992) *Questionnaire design, interviewing and attitude measurement, New ed*. New York, Pinter Publishers.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P. and Moher, D. (2021) 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', *BMJ*, 372, p. n71.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P. and Dierolf, U. (2013) 'Making Research Data Repositories Visible: The re3data.org Registry', *PLOS ONE*, 8(11), p. e78080.

Patil, S., Lu, H., Saunders, C.L., Potoglou, D. and Robinson, N. (2016) 'Public preferences for electronic health data storage, access, and sharing — evidence from a pan-European survey', *Journal of the American Medical Informatics Association*, 23(6), pp. 1096-1106.

Pearce, M.S., Unwin, N.C., Parker, L. and Craft, A.W. (2009) 'Cohort Profile: The Newcastle Thousand Families 1947 Birth Cohort', *International Journal of Epidemiology*, 38(4), pp. 932-937.

Pereira, S., Gibbs, R.A. and McGuire, A.L. (2014) 'Open access data sharing in genomic research', *Genes*, 5(3), pp. 739-747.

Petticrew, M. and Roberts, H. (2006) *Systematic Reviews in the Social Sciences: a practical guide*. Oxford: Blackwell Publishing Ltd.

Pisani, E., Whitworth, J., Zaba, B. and Abou-Zahr, C. (2010) 'Time for fair trade in research data', *Lancet*, 375(9716), pp. 703-5.

Platt, J. and Kardia, S. (2015) 'Public Trust in Health Information Sharing: Implications for Biobanking and Electronic Health Record Systems', *Journal of Personalized Medicine*, 5(1), pp. 3-21.

Platt, J.E., Jacobson, P.D. and Kardia, S.L.R. (2017) 'Public Trust in Health Information Sharing: A Measure of System Trust', *Health Services Research*, 53(2), pp. 824-845.

PLOS. (2014) *Data Availability*. Available at: http://journals.plos.org/plosone/s/data-availability (Accessed: 08/12/2021).

Polanin, J.R. and Terzian, M. (2019) 'A data-sharing agreement helps to increase researchers' willingness to share primary data: results from a randomized controlled trial', *J Clin Epidemiol*, 106, pp. 60-69.

PRIME Centre Wales (2018) *SUPER group*. Available at: http://www.primecentre.wales/super-group-update.php (Accessed: 08/12/2021).

Prisco, D., Ciuti, G., Grifoni, E., Silvestri, E. and Emmi, G. (2016) 'Sharing data of clinical trials', *European Journal of Internal Medicine*, 33, pp. e25-e26.

PROSPERO (2017) *international prospective register of systematic reviews,* Available at: https://www.crd.york.ac.uk/PROSPERO/#aboutpage (Accessed: 08/12/2021).

QualtricsXM (2021) *Experience Design Experience Improvement*. Available at: https://www.qualtrics.com/uk/ (Accessed: 08/12/2021).

re3data.org (2020) *Registry of Research Data Repositories*. Available at: https://doi.org/10.17616/R3D (Accessed: 08/12/2021).

Redish, J. (2000) 'Readability formulas have even more limitations than Klare discusses', *ACM J. Comput Doc*, 24(3), pp. 132–137.

Richards, S.H., Campbell, J.L., Walshaw, E., Dickens, A. and Greco, M. (2009) 'A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires', *Medical Education*, 43(8), pp. 757-766.

Robling, M., Hood, K., Houston, H., Pill, R., Fay, J. and Evans, H. (2004) 'Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study', *Journal of Medical Ethics*, 30(1), pp. 104-109.

Rogith, D., Yusuf, R.A., Hovick, S.R., Peterson, S.K., Burton-Chase, A.M., Li, Y., Meric-Bernstam, F. and Bernstam, E.V. (2014) 'Attitudes regarding privacy of genomic information in personalized cancer therapy', *J Am Med Inform Assoc*, 21(e2), pp. e320-5.

Ross, J.S. and Krumholz, H.M. (2013) 'Ushering in a new era of open science through data sharing: The wall must come down', *JAMA*, 309(13), pp. 1355-1356.

SAIL Databank (2020) *Public Engagement*. Available at: https://saildatabank.com/about-us/public-engagement/ (Accessed: 08/12/2021).

Savitz, D.A. and Olshan, A.F. (1998) 'Describing Data Requires No Adjustment for Multiple Comparisons: A Reply from Savitz and Olshan', *American Journal of Epidemiology*, 147(9), pp. 813-814.

Schonlau, M., Fricker, R.D. and Elliott, M.N. (2002) *Conducting Research Surveys via E-mail and the Web*, California, RAND Corporation.

Scott, B. (2017) *Text Readability Consensus Calculator*. Available at: https://readabilityformulas.com/free-readability-formula-tests.php (Accessed: 08/12/2021).

ScottisH Informatics Programme (SHIP) (2013) *The collation, management, dissemination and research analysis of anonymised Electronic Patient Records,* Available at: http://www.scot-ship.ac.uk/contact.html (Accessed: 08/12/2021).

Shabani, M., Bezuidenhout, L. and Borry, P. (2014) 'Attitudes of research participants and the general public towards genomic data sharing: a systematic literature review', *Expert Review of Molecular Diagnostics*, 14(8), pp. 1053-1065.

Shabani, M. and Obasa, M. (2019) 'Transparency and objectivity in governance of clinical trials data sharing: Current practices and approaches', *Clinical Trials*, p. 1740774519865517.

Shah, N., Coathup, V., Teare, H., Forgie, I., Giordano, G.N., Hansen, T.H., Groeneveld, L., Hudson, M., Pearson, E., Ruetten, H. and Kaye, J. (2018) 'Sharing data for future research—engaging participants' views about data governance beyond the original project: a DIRECT Study', *Genetics in Medic*ine, 21(5), pp. 1131-1138.

Shorten, A. and Smith, J. (2017a) 'Mixed methods research: expanding the evidence base', *Evidence Based Nursing*, 20(3), pp. 74-75.

Sim J. (1998) 'Collecting and analysing qualitative data: issues raised by the focus group', *Journal of Advanced Nursing,* 28(2), pp. 345-52.

Skelly, A.C. (2011) 'Probability, proof, and clinical significance', *Evidence-based spine-care journal*, 2(4), pp. 9-11.

Smithson, J. (2000) 'Using and analysing focus groups: Limitations and possibilities', *International Journal of Social Research Methodology*, 3(2), pp. 103-119

Stettler, K. and Featherston, F. (2012) 'Early Stage Scoping: Bridging the Gap between Survey Concepts and Survey Questions', *Fourth International Conference on Establishment Statistics*. Montreal, Canada.

Stone, M.A., Redsell, S.A., Ling, J.T. and Hay, A.D. (2005) 'Sharing patient data: competing demands of privacy, trust and research in primary care', *British Journal of General Practice*, 55(519), pp. 783-789.

Strycharz, J., Ausloos, J. and Helberger, N. (2020) 'Data Protection or Data Frustration? Individual Perceptions and Attitudes Towards the GDPR', *European Data Protection Law Review*, 6(3), pp. 407-421.

Sturges, P., Bamkin, M., Anders, J.H.S., Hubbard, B., Hussain, A. and Heeley, M. (2015) 'Research data sharing: Developing a stakeholder-driven model for journal policies', *Journal of the Association for Information Science and Technology*, 66(12), pp. 2445-2455.

Sullivan, G.M. and Artino, A.R., Jr. (2013) 'Analyzing and interpreting data from likert-type scales', *Journal of graduate medical education*, 5(4), pp. 541-542.

Sydes, M.R., Johnson, A.L., Meredith, S.K., Rauchenberger, M., South, A. and Parmar, M.K. (2015) 'Sharing data from clinical trials: the rationale for a controlled access approach', *Trials*, 16(1), pp. 104-110.

Taichman, D.B., Backus, J., Baethge, C., Bauchner, H., de Leeuw, P.W., Drazen, J.M., Fletcher, J., Frizelle, F.A., Groves, T., Haileamlak, A., James, A., Laine, C., Peiperl, L., Pinborg, A., Sahni, P. and Wu, S. (2016) 'Sharing Clinical Trial Data: A Proposal from the International Committee of Medical Journal Editors', *PLoS Med*, 13(1), pp. 41-43.

Tashakkori, A. and Teddlie, C. (2010) 'Realism as a Stance for Mixed Methods Research,' in *SAGE Handbook of Mixed Methods in Social &; Behavioral Research*. Available at: https://methods.sagepub.com/book/sage-handbook-of-mixed-methods-social-behavioral-research-2e (Accessed: 08/12/2021).

Taylor, M.J. and Taylor, N. (2014) 'Health research access to personal confidential data in England and Wales: assessing any gap in public attitude between preferable and acceptable models of consent', *Life Sciences, Society and Policy*, 10(1), pp. 15-39.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. and Frame, M. (2011) 'Data Sharing by Scientists: Practices and Perceptions', *PLOS ONE*, 6(6), p. e21101.

The Academy of Medical Sciences (2013) *Clinical trials data sharing: science, privacy and ethics*: Note of the dinner discussion, 28 November 2013, Available at: https://acmedsci.ac.uk/publications (Accessed 08/12/2021).

The Academy of Medical Sciences (2016) *Summary of a joint workshop to explore the ICMJE proposal on 'Sharing clinical trial data'*. https://acmedsci.ac.uk/file-download/41607-58206ac524230.pdf (Accessed 08/12/2021).

The Wellcome Trust (2010) *Policy on Data Management and Sharing, Available at: https://www.nature.com/articles/npre.2011.6007.1* (Accessed: 08/12/2021).

Thomas, J. and Harden, A. (2008) 'Methods for the thematic synthesis of qualitative research in systematic reviews', *BMC Medical Research Methodology*, 8(1), pp. 1-10.

Thwaites Bee, D. and Murdoch-Eaton, D. (2016) 'Questionnaire design: the good, the bad and the pitfalls', *Arch Dis Child Educ Pract Ed*, 101(4), pp. 210-2.

Transcribe.com (2015) *How to Transcribe a Focus Group the Right Way,* [online]. Available at: https://www.transcribe.com/how-to-transcribe-a-focus-group-the-right-way/ (Accessed: 08/12/2021).

Treweek, S., Doney, A. and Leiman, D. (2009) 'Public attitudes to the storage of blood left over from routine general practice tests and its use in research', *Journal of Health Services Research & Policy*, 14(1), pp. 13-19.

Trinidad, S.B., Fullerton, S.M., Bares, J.M., Jarvik, G.P., Larson, E.B. and Burke, W. (2010) 'Genomic research and wide data sharing: Views of prospective participants', *Genet Med*, 12(8), pp. 486-495.

Tucker, K., Branson, J., Dilleen, M., Hollis, S., Loughlin, P., Nixon, M.J. and Williams, Z. (2016) 'Protecting patient privacy when sharing patient-level data from clinical trials', *BMC Medical Research Methodology*, 16(1), pp. 77-87.

Tudur-Smith, C, Hopkins, C, Sydes, M, Woolfall, K, Clarke, M, Murray, G and Williamson, P (2015) *Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials* (Version 1). MRC Network of Hubs for Trials Methodology, MRC Network of Hubs for Trials Methodology.

Tudur-Smith, C., Dwan, K., Altman, D.G., Clarke, M., Riley, R. and Williamson, P.R. (2014) 'Sharing Individual Participant Data from Clinical Trials: An Opinion Survey Regarding the Establishment of a Central Repository', *PLOS ONE*, 9(5), p. e97886.

UCLA Advanced Research Computing (2021) HOW CAN I DO POST-HOC PAIRWISE COMPARISONS USING STATA? Available at: https://stats.oarc.ucla.edu/stata/faq/faqhow-can-i-do-post-hoc-pairwise-comparisons-using-stata/ (Accessed: 09/06/2022).

UK Data Archive (2015) *The UK's Largest Collection of Digital Research Data in the Social Sciences and Humanities*. Available at: www.data-archive.ac.uk (Accessed: 08/12/2021).

UK Data Service (2016) *Manage data- plan to share*. Available at: https://www.ukdataservice.ac.uk/deposit-data/how-to (Accessed: 08/12/2021).

UK Data Service (2020) *2011 UK Townsend Deprivation Scores*. Available at: https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-uk-townsend-deprivation-scores (Accessed: 08/12/2021).

UK Data Service (2021) *Data access policy*. Available at: https://www.ukdataservice.ac.uk/get-data/data-access-policy/controlled-data.aspx (Accessed: 08/12/2021).

UK Parliament (1998) *Data Protection Act 1998*. TSO (The Stationary Office).

UK Research and Innovation (2018) *Guidance on best practice in the management of research data*. Available at: https://www.ukri.org/files/legacy/documents/rcukcommonprinciplesondatapolicy-pdf/ (Accessed: 08/12/2021).

UK Research and Innovation (2021) *Co-production in research*. Available at: https://www.ukri.org/about-us/policies-standards-and-data/good-research-resource-hub/research-co-production/ (Accessed: 08/12/2021).

UKCRC (2021) *UKCRC Registered Clinical Trials Units Network: About us*. Available at: https://www.ukcrc-

ctu.org.uk/page/about#:~:text=The%20UKCRC%20Registered%20CTU%20Network%20is%20a%20network,consists%20of%20CTU%20members%20from%20across%20the%20UK. (Accessed: 08/12/2021).

Understanding Patient Data (2017) *New words and pictures to explain anonymisation*. Available at: https://understandingpatientdata.org.uk/news/new-words-and-pictures-explain-anonymisation (Accessed: 08/12/2021).

Vallance, P., Freeman, A. and Stewart, M. (2016) 'Data Sharing as Part of the Normal Scientific Process: A View from the Pharmaceutical Industry', *PLoS Med*, 13(1), p. e1001936.

van Panhuis, W.G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A.J., Heymann, D. and Burke, D.S. (2014) 'A systematic review of barriers to data sharing in public health', *BMC Public Health*, 14(1), pp. 1144-1153.

Vehovar, V. and Beullens, K. (2018) 'Cross-National Issues in Response Rates', in David L. Vannette, J.A.K. (ed.) *The Palgrave Handbook of Survey Research*. Springer ebooks: London, Palgrave Macmillan.

Vickers, A.J. (2006) 'Whose data set is it anyway? Sharing raw data from randomized trials', *Trials*, 7(1), pp. 15-21.

Villar, A. (2011) 'Response Bias', in Lavrakas, P.J. (ed.) *Encyclopedia of Survey Research Methods* Thousand Oaks, California: Sage Publications.

Vivli (2021) *Independent Review Panel*. Available at: https://vivli.org/about/independent-review-panel/ (Accessed: 29/07/2021).

Vlahou, A., Hallinan, D., Apweiler, R., Argiles, A., Beige, J., Benigni, A., Bischoff, R., Black, P.C., Boehm, F., Céraline, J., Chrousos, G.P., Delles, C., Evenepoel, P., Fridolin, I., Glorieux, G., Gool, A.J.v., Heidegger, I., Ioannidis, J.P.A., Jankowski, J., Jankowski, V., Jeronimo, C., Kamat, A.M., Masereeuw, R., Mayer, G., Mischak, H., Ortiz, A., Remuzzi, G., Rossing, P., Schanstra, J.P., Schmitz-Dräger, B.J., Spasovski, G., Staessen, J.A., Stamatialis, D., Stenvinkel, P., Wanner, C., Williams, S.B., Zannad, F., Zoccali, C. and Vanholder, R. (2021) 'Data Sharing Under the General Data Protection Regulation', *Hypertension*, 77(4), pp. 1029-1035.

VOICE (2017) *What is Voice North*? Available at: https://www.voice-global.org/latest/2016/november-2016/directors-blog-what-is-voice-north/ (Accessed: 08/12/2021).

Walport, M. and Brest, P. (2011) 'Sharing research data to improve public health', *The Lancet*, 377(9765), pp. 537-539.

Whyte, A. (2015) *'Where to keep research data: DCC checklist for evaluating data repositories' v.1.1*. Available at: https://www.dcc.ac.uk/guidance/how-guides/where-keep-research-data (Accessed: 08/12/2021).

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3(1), p. 160018.

Wilkinson, S. (1998) 'Focus group methodology: a review', *International Journal of Social Research Methodology*, 1(3), pp. 181-203.

Williams, G. and Pigeot, I. (2017), 'Consent and confidentiality in the light of recent demands for data sharing', *Biometrical Journal*, 59(2), pp. 240-250.

Williamson, P.R., Altman, D.G., Bagley, H., Barnes, K.L., Blazeby, J.M., Brookes, S.T., Clarke, M., Gargon, E., Gorst, S., Harman, N., Kirkham, J.J., McNair, A., Prinsen, C.A.C., Schmitt, J., Terwee, C.B. and Young, B. (2017) 'The COMET Handbook: version 1.0', *Trials*, 18(3), pp. 280-330.

Willis, G. (2005) *Cognitive Interviewing: A Tool For Improving Questionnaire Design* [online], Available at: https://methods.sagepub.com/book/cognitive-interviewing, (Accessed 08/12/2021).

Willis, G.B. (1999) *Cognitive Interviewing a "how to" guide* [online]. National Center for Health Statistics, Available at: https://www.hkr.se/contentassets/9ed7b1b3997e4bf4baa8d4eceed5cd87/gordonwillis.pdf (Accessed 08/12/2021).

Willis, G.B. and Lessler, J.T. (1999) *Question Appraisal System QAS-99, A guide for systematically evaluating survey question wording*. Rockville: Research Triangle Institute.

Willison, D.J., Steeves, V., Charles, C., Schwartz, L., Ranford, J., Agarwal, G., Cheng, J. and Thabane, L. (2009) 'Consent for use of personal information for health research: do people with potentially stigmatizing health conditions and the general public differ in their opinions?', *BMC Med Ethics*, 10(10).

Xafis, V. The acceptability of conducting data linkage research without obtaining consent: lay people's views and justifications. *BMC Med Ethics,* 16, 79 (2015).

Appendices

Review Article

Journal of Health Services Research & Policy
2018, Vol. 23(2) 123–133

DOI: 10.1177/1355819617751555
journals.sagepub.com/home/hsr
SAGE

Systematic review of participants' attitudes towards data sharing: a thematic synthesis

Nicola Howe[1], Emma Giles[2], Dorothy Newbury-Birch[3] and Elaine McColl[4]

Abstract

Objectives: Data sharing is well established in biological research, but evidence on sharing of clinical trial or public health research study data remains limited, in particular studies of research participants' perspectives of data sharing. This study systematically reviewed international evidence of research participants' attitudes towards the sharing of data for secondary research use.

Methods: Systematic search of seven databases, and author-, citation- and bibliography-follow up to identify studies examining research participants' attitudes towards data sharing. Studies were thematically analysed using NVivo v10 to identify recurring themes.

Results: Nine studies were eligible for inclusion. Thematic analysis identified four key themes: (1) benefits of data sharing, including benefit to participants or immediate community, benefits to the public and benefits to science or research; (2) fears and harms, such as fear of exploitation, stigmatization or repercussions, alongside concerns about confidentiality and misuse of data; (3) data sharing processes, in particular the role of consent in the process; and (4) the relationship between participants and research such as trust in different types of research or organization and the relationship with the original research team.

Conclusions: The available literature on attitudes towards sharing data from clinical trials or public health interventions remains scant. This study has identified four themes regarding research participants' attitudes and preferences, which should be considered by policy makers, and explored with further research.

Keywords

attitude, data sharing, research participant

Background

In 2011, a '*joint statement of purpose*'[1] from global health funding agencies, academic researchers, international organizations and journals promoted data sharing in health research. Similar statements have followed since, such as those by the International Committee of Medical Journal Editors (ICMJE)[2] and the US Institute of Medicine (IOM).[3] These highlight the '*ethical obligation to share*'[2] and they encourage a culture where '*data sharing is the expected norm*'.[3] Many journals [1,2,4,5] now require research data to be shared after study completion, for example, through a recognized repository.[6]

Pooling data or conducting secondary analysis of data already collected for another study is expected to accelerate the '*pace of discovery*',[1] advance science and clinical knowledge[3] and identify safe and effective patient treatments more quickly. Sharing also allows independent confirmation of results[2] and minimizes

[1]Database Manager, Newcastle Clinical Trials Unit, Newcastle University, UK
[2]Senior Research Lecturer in Public Health, School of Health and Social Care, Teesside University, UK
[3]Professor of Alcohol and Public Health Research, School of Health and Social Care, Teesside University, UK
[4]Professor of Health Service Research, Institute of Health and Society, Newcastle University, UK

Corresponding author:
Nicola Howe, Clinical Trials Unit, Newcastle University, 1–4 Claremont Terrace, Newcastle upon Tyne NE2 4AE, UK.
Email: nicola.howe@ncl.ac.uk

repetition of research and so reduces associated costs. It encourages transparency and reproducibility, so increasing the overall quality of research.[7] The value of participants' participation is maximized[1] by '*potentially facilitating additional findings beyond the original. . .outcomes*'.[3] Overall, data sharing increases value for money for funders, while both honouring the contribution that participants made and fulfilling researchers' moral obligation to participants, who may have put their health at risk to take part in research.[2]

However, although there is a well-established culture of data sharing in the genetic and genomic communities, data sharing is less ingrained in public health research.[1] To increase the potential for sharing, a combination of gaining consent, anonymizing and regulating access is needed.[8] Much attention has been paid to anonymization and data control, but it is unclear how well suited the consent process is to sharing.

There is a limited amount of literature on perspectives of research participants on data sharing, particularly those participating in clinical trials or public health interventions. Available work tends to focus on primary care, (electronic) health records or biobank data,[9–11] where participants are accepting of their data to be used in clinical studies, but it should be anonymized with consent sought in advance.[9,10–13] Evidence on genetic and health record data further suggests that participants prefer to be contacted before their data are used in subsequent research.[14–16] They may also want to re-consent based on the type of secondary research to be conducted, distinguishing between 'acceptable' (e.g. health service) and 'unacceptable' (e.g. commercial) research.[9,17–19]

This study reviews the international literature on research participants' attitudes towards data sharing in the context of clinical trials and other public health research. It specifically explores participants' understanding of data sharing, their attitudes towards sharing and whether awareness of data sharing could affect consent to take part in research.

Methods

A protocol was developed using the PRISMA-P, 2015 checklist[20] and followed throughout the systematic review process (online Appendix 1). The protocol was not eligible for PROSPERO registration as it does not concern health outcomes.[21]

Search strategy

We piloted search terms in a Medline scoping search and then used broad search terms (online Appendix 2) relating to data sharing and participant, patient or public attitudes to interrogate the following databases: Medline, Embase, Web of Science, ASSIA, CINAHL, HMIC and PsychINFO. Key terms were taken from studies already identified and adapted for each database. Letters to the editor, books, conference proceedings and editorials were excluded. Reference and citation lists of included studies, publications of included first authors and references of systematic reviews were also searched; systematic reviews as such were excluded.

Inclusion and exclusion criteria

To be included studies had to report qualitative, quantitative or mixed methods empirical research. They had to address data sharing regarding secondary use of research data already collected as part of a trial, study or intervention. Included studies further had to examine attitudes of research participants or potential participants, i.e. members of the public. They had to be published between 1995 (year of publication of EU Directive 95/46/EC; the Data Protection Directive)[22] and 25 January 2017.

Studies concerning biobank data, human tissue, blood samples, routinely collected primary and secondary care data (health records) or 'data-linkage' were excluded.

We set no restrictions on language or country of origin.

Study selection

One reviewer (NH) screened all titles and a second reviewer (DNB) independently screened 20%, erring towards inclusion if uncertain. The same reviewers then independently screened all accepted abstracts against the inclusion criteria, noting reasons for exclusion. Reconciliation of disagreement was achieved through discussion and by erring towards inclusion. The remaining full papers were read in detail (by NH), with later group discussion including three authors (DNB, EM, NH) (Figure 1).

Data extraction and quality appraisal

Detailed data were extracted from each included study by one author (NH) according to country of origin, date of research, study design, participant characteristics, study aims and key themes identified by authors of included papers (Table 1).

Each included study was assessed for quality by two reviewers (NH and either EM or DNB), using the Critical Appraisal Skills Framework Qualitative Appraisal Tool (CASP)[23] for qualitative studies[24–31] and the Best Bets Survey Checklist Quality Assessment Tool[32] for the study using quantitative methods.[33]

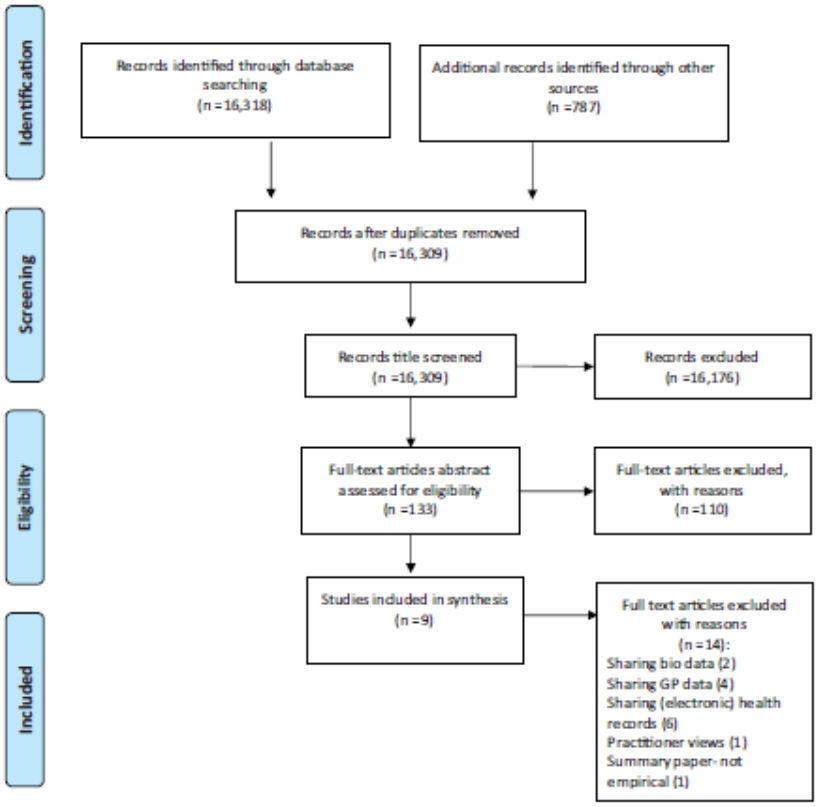


Figure 1. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) flow diagram.

Data synthesis and analysis

Results sections from included studies were analysed using thematic synthesis.[34] This was done in NVivo Software Version 10 by one author (NH) using a line-by-line approach, inductively highlighting all relevant quotes from participants or descriptions from the authors of the original studies to form 'free codes'. We did not consider quotes and data from non-participants (e.g. where researchers were also interviewed) or any corresponding author description. It was possible that the same sentence was assigned more than one code. The discussion section of the sole quantitative paper[33] was also analysed to provide a greater richness of descriptive data.

Free codes thus derived were then amalgamated into descriptive groups by two reviewers (NH and ELG), using a hierarchical structure. This process was repeated until the groups became broad themes, which were interrogated by all authors. Grouping codes in NVivo was an evaluative process; the original text was referred to ensuring that codes were not taken out of their intended context. There was no attempt to produce analytical themes,[34] as the purpose of this review was simply to report emerging themes, not to speculate as to why they occurred.

Results

Description of included studies

Of the 16,309 records identified by searches, nine met the inclusion criteria (Figure 1). The studies were published between 2002 and 2016 originating from Japan, Thailand, India, Kenya, Canada, Vietnam and the USA. All but one study used qualitative methods, such as focus groups or interviews. The remaining study[33] quantitatively analysed a telephone survey. Five studies concerned data sharing in low- and middle-income countries. These were part of the same funding award and shared many of the same authors and employed common methods.

Table 1. Characteristics of nine included studies addressing participant attitudes towards data sharing.

Study	Country of origin	Study design	Participant characteristics	Aim	Key themes of study	Quality appraisal
Asai et al.[24]	Japan	Focus group interviews and brief demographic questionnaire.	Lay participants aged 35–55, married with children, with experience or relatives' experience of inpatient care during the preceding five years. No close family members who were health care professionals.	To explore lay persons' attitudes toward the use of archived (existing) materials such as medical records and biological samples (and to compare them with the attitudes of physicians who are involved in medical research).	• Types of consent • Prerequisites for sharing • Benefits to public • Ownership of medical records • Trust in researchers	CASP: 8/10 'yes' answers
Cheah et al.[25]	Thailand	Focus group with seven, interview with one. Topic guides taken from a template developed collaboratively with partners from other sites.	Community members acting as 'community representatives', affiliated with Shoklo Malaria Research Unit where they had been hired as temporary community engagement staff.	To understand attitudes and experiences of relevant stakeholders about what constitutes good data sharing practice.	• Benefits of sharing • Concerns and harms • Suggestions for best practice	CASP: 9/10 'yes' answers
Hate et al.[26]	India	Focus groups conducted at outreach centres. Attended by field workers as a reassuring presence. Series of scenarios presented that drew on previous contributions to research.	(Employees or) participants in research conducted by SNEHA. Participants were familiar with the organization and its work. 20 female community members.	To identify features of ethical data sharing practice in the context of research involving women and children in informal settlements. Specific objectives were to examine stakeholders' understandings, concerns, and hopes about what would happen to data and their views on what might constitute good data sharing practice; to identify models of data sharing and governance currently in use; to examine contextual considerations affecting data sharing processes; to	• Benefits of data sharing • Harms of sharing • Barriers to sharing • Obligations and responsibilities • Prerequisites for data sharing • Governance and policy • Broad, middle and explicit consent.	CASP: 10/10 'yes' answers

(continued)

Table 1. Continued

Study	Country of origin	Study design	Participant characteristics	Aim	Key themes of study	Quality appraisal
				identify perceived principles of good practice in data sharing and to consider suitable methods of developing appropriate data sharing processes.		
Jao et al.[27]	Kenya	Small group discussions (5–6 people) lasting 3–4 hours. After discussion groups, 3–4 individuals were chosen (reflecting differences in attitude and gender) for interviews lasting 30–45 min.	A range of stakeholders including 30 community members including assistant chiefs (6) and community representatives (24) with relatively low research experience.	A consultation on data sharing, mapping the views and values of diverse stakeholders in a large international research program, the Kenya Medical Research Institute (KEMRI). This paper focuses on views on 'fair processes' in data sharing.	• Types of consent • Informed consent process • Community engagement • Feedback on data sharing process • Oversight for decisions on access to data • Perceived benefits and challenges	CASP: 10/10 'yes' answers
Jao et al.[28]	Kenya	Small group discussions (4–6 people) with case study and vignette. Emerging findings noted and used to prompt discussion. After discussion groups, 3–4 individuals were chosen (reflecting differences in attitude and gender) for interviews lasting 30–45 min.	Community representatives: 'typical' community members selected by and from local villages at public meetings to support interactivity for a three-year period, and participate in annual workshops on research-related topics.	To report research stakeholders perceptions of benefits and challenges in sharing data and the emerging importance of trust at individual and institutional levels.	• Importance of data sharing • Challenges and concerns for primary communities • Risks of harms • Fairness to the primary community • Challenges and harms for originating researchers • Misuse of data • Does it matter who's asking?	CASP: 10/10 'yes' answers
Manhas et al.[29]	Canada	Semi-structured interview guide used in focus groups and	Maternal and paternal participants in two longitudinal	To explore parent perspectives about sharing their	• Altruism has limits • Participants have	CASP: 10/10'yes' answers

(continued)

Table 1. Continued

Study	Country of origin	Study design	Participant characteristics	Aim	Key themes of study	Quality appraisal
		individual interviews. Recruitment, data collection and analysis continued until data saturation reached.	pregnancy cohort research studies. Purposive sampling to identify participants who were fathers and mothers, older and younger than 30, visible minorities and new immigrants.	own, and their child's non-biological data.	ongoing privacy concerns • Some participants believe that congruence in values between themselves and research/ researchers is important	
Manhas et al.[30]	Canada	Group and individual interviews	Maternal and paternal participants in two longitudinal pregnancy cohort research studies. Purposive sampling to identify participants who were fathers and mothers, older and younger than 30, visible minorities and new immigrants.	To examine parent preferences for sharing non-biological data, specifically in regards to the consent process.	• Reciprocity: parents want reciprocity among participants, repositories and researchers regarding respect and trust. • Accuracy: parents worry about the interrelationships between validity of the consent processes and secondary data use.	CASP: 10/10 'yes' answers
Merson et al.[31]	Vietnam	Focus groups with participants and their families.	15 Clinical research participants enrolled in observational or cohort studies from northern and southern, rural and urban centres.	To explore stakeholders' understanding, perceptions, experiences attitudes and concerns about sharing individual level clinical data.	• Views about a novel initiative • Views about acceptable sharing • Trust • Consent	CASP: 9/10 'yes' answers
Platt and Kardia[33]	USA	119-item survey developed to evaluate predictors of trust in the health	447 Members of the general public 51.5% male, aged 18–65 (most aged	To identify characteristics of the general public that predict trust in a health system that includes	• Knowledge of health information sharing • Privacy concerns	Best Bets Survey Checklist Quality Assessment Tool: Paper rating 7/10

(continued)

Table 1. Continued

Study	Country of origin	Study design	Participant characteristics	Aim	Key themes of study	Quality appraisal
		system, broadly defined as a web of relationships among health care providers, departments of health, insurance systems and researchers. Included six trust characteristics included in conceptual model as well as additional questions about trust in specific institutions.	26–34). White (76.1%) Black (7.16%), Asian (8.05%), Hispanic (4.70%), other (3.13%). Most were college or some college educated. 62% non-home owners. Self-rated health, excellent 18%, very good 40% good 29%, fair 11%, poor 1.6%.	researchers, health care providers, insurance companies and public health departments. Regarding Data Sharing in particular: 'our study looks to see whether knowledge impacts trust in data sharing and if so, whether or not it increases support'.	• Expectations of benefit	

CASP: critical appraisal skills framework qualitative appraisal tool.

Quality appraisal

All studies scored highly on the quality appraisal (Table 1). However, three qualitative studies lacked detail about the relationship between researcher and participant.[24–26,31] The study by Platt and Kardia[33] did not explicitly state sample size, response rate and non-responders.

Themes arising from qualitative analysis

Analysis of the studies identified four themes: (1) benefits of data sharing, (2) fears and harms, (3) data sharing processes and (4) relationship between participants and research. We examine each in turn.

Benefits of data sharing

In all studies, participants identified benefits of data sharing, with three main types emerging: benefit to participants or immediate community, benefits to the public, and benefits to science or research.

Most participants wanted to see the benefits of data sharing in their local community, with one participant summarizing: '*Data sharing is acceptable if the community benefits…; there is no point in merely writing about issues*'.[26] There should be '*local translational benefits*'[28] for '*the community that contributed*',[25] particularly if the research in question focussed on a burden the community faced.[26]

The 'expectation of benefit'[33] from data sharing also extended to the wider public, with phrases such as '*greater good*',[29] '*social value*'[25] and '*actually helping people*'[30] used in one form or another by research participants. Jao et al.[28] reported that public benefit was sometimes seen as '*satisfied by the involvement of international institutions… such as the World Health Organization*', suggesting that the perception of benefit may be as important as actually experiencing it.

Participants also appreciated the benefits to science and research, explaining that sharing '*increased the efficiency of research and researcher opportunities*',[29] '*generated evidence*' and '*avoided duplication of effort*'.[26] Participants thought that local researchers should also benefit, and that their '*careers should not be "overtaken" by others who had made less investment*'.[28]

Fears and harms

Participants expressed fear of exploitation, stigmatization or repercussions, with some mentioning specific '*harms*'[33] such as being reported to social services or an attempted abduction of their child,[29] alongside more mundane concerns such as third-party contact or telemarketing. Some participants reported that

they would be hesitant to share their data as they were sceptical that it would be used in the right way, and so were likely to consent somewhat '*reluctantly*'.[24,29,31]

Participants wanted to maintain an element of control of their data, highlighting feelings of powerlessness, as there was '*no way for us to know whether or not our personal information is dealt with anonymously*'.[24] '*Personal information*' was described as '*something that can let people know who you are*'.[29] Concern about being identifiable or the desire for privacy/confidentiality was referred to in most of the included studies.[24–26,28,29,31] Some participants talked about the distinction between '*sensitive*' (e.g. personal details, ethnicity,[25] HIV status, history of abuse[26]) and less sensitive data such as routine demographics.[27] The potential sensitivity of data was, however, related to its intended use.[27]

Participants were concerned that data could be '*misused*',[26,28,29,31] either unintentionally (misinterpretation) or deliberately, in order to contact participants or manipulate data to suit a particular purpose. '*Misuse*' was therefore about both confidentiality and about aligning secondary research with participants' principles. Harm was considered more likely to occur if data were shared '*outside the original research team*',[25] with participants worrying about identification if data were used in ways not initially anticipated. One participant reflected on the need for '*penalties*' for secondary researchers if their data were used in ways '*not affiliated*' with the original research: '*. . .if they use it for personal gain or a third party company. . .*'.[29]

Data sharing processes

Identified barriers to data sharing included their '*novelty*',[27] '*limited precedent*'[26] and practicalities such as the time or work involved to prepare data for sharing.[25–27,29,31] Participants recognized the resources required to implement data sharing, with phrases such as '*resource implications*', '*funding and capacity building*'[25] and '*substantial work*'[26] given by authors to paraphrase participants' views.

Studies based in low- and middle-income countries[25–28,31] specifically emphasized community or stakeholder involvement, while participants' desire to be involved in the data sharing process was identified in all studies, as was the desire to be notified when their data were (re)used and to be informed of the results of studies using their data.

Participants showed varying degrees of understanding of the consent process. Some participants saw the consent process as an informative tool that can play a '*wider educational role*'.[27] '*Perhaps you can explain in the consent form. . . other researchers can access my data to do further research*'.[31]

Seven studies[24–27,29–31] discussed different levels of consent with participants. For some, a broad initial consent would be acceptable, while others wished for '*individual informed consent*' or '*personal permission*'.[24] Participants evaluated the practicalities of each approach but stated their preference based on ideals of respect and transparency: '*we always like to be asked. . .I don't think [the project-specific consent model is] a great idea, but I think it would make us feel good*'.[30] For others, it depended on whom the data would be shared with, and they would evaluate on a '*case-by-case basis*'.[25]

Re-consenting was described in one study as an '*unnecessary inconvenience*'[27] and an '*annoyance*' or '*irritation*',[30] which risked inviting more questions than if researchers had just shared data anyway. References were made to the practical difficulty of re-consenting participants.[25,27,30,31]

To be more comfortable with data sharing, participants wanted better data governance or gatekeepers, with processes to store data and manage access requests.[25–27,29–31] Research data repositories (RDRs) could act as '*stewards*' for data[29] perhaps with a committee who could oversee data sharing requests.[26,27,29,30] A committee would be '*a group trusted to make decisions*',[27] ideally with lay representatives, who could reach a consensus, and be held accountable for sharing decisions.[29]

Participants identified other conditions that they would like to see in place before they could comfortably agree to share their data, including participants having understood that their data could be shared (transparency), risks mitigated, the research being in the publics' interest and the research being congruent with the participants' values.[24–27,29,31]

Relationship between participants and research

Some studies reported that participants were largely unaware that researchers might already be sharing their data.[24,26–30] Participants wanted data sharing to be better publicized, or be given the option to choose whether or not to share. Where participants were subsequently informed about data sharing, it was largely accepted as a '*necessary sacrifice*' for scientific or medical progress.[29] There were then '*high levels of uncertainty about how data might be used once it had been shared*',[28] with a desire for transparency regarding the recipient's intentions.

The idea that their data could be shared with a secondary researcher prompted participants to consider acceptable types of research or researcher. Although one participant was content with anyone '*[a]s long as it's a qualified researcher*',[29] others wanted information about the researchers before agreeing that their data

could be shared,[24] based on the idea that you '*...approve secondary researchers, not their projects*'.[29]

Most participants agreed that their data should not be used for commercial gain. '*[T]hird parties*'[27,29] or '*industry based researchers*'[29] were distrusted because they might use data in a way that is inconsistent with the values of the participant or attempt to contact them '*for nefarious or unconsented purposes*'[29] (e.g. telemarketing). One participant stated that if their data were to be shared with '*a for-profit research group or something, I would want to know and at that point I would actually probably opt out*'.[29]

If participants allowed their data to be shared, researchers should ensure that they make good or proper use of it.[24,29] It would be '*wrong*' to use the data in a way that the participant is unlikely to have agreed to or understood.[27,28] Participants were placing a great deal of trust in researchers to share their data with appropriate collaborators.

The researcher–participant relationship was described as '*socially unequal*', a '*tacit agreement between the researchers and patients*'[24] and similar to the '*patient-provider*' relationship,[33] with the researchers indebted to participants.[26] The relationship with the originating researcher was crucial because it was they who would inform, reassure and garner a willingness to share. The primary researcher was also the preferred point of contact regarding re-consent '*...just you guys*'.[30]

Discussion

Previous reviews have explored participants' attitudes towards the sharing of biological and health record data, or data linkage.[11,35,36,37] Our review identifies similar concerns: participants are open to and understand the advantages of data sharing, but they lack awareness and have concerns regarding confidentiality, potential data misuse, governance and commercial data use.

Participants in the included studies wanted appropriate data protection, and they identified processes that they thought could be modified in order to promote acceptance of data sharing. There was less evidence regarding the effects of data sharing on agreement to participate in research in the first place.

Implications for research and practice

In the current global drive to accommodate data sharing from the outset of studies,[1–5] this review provides evidence of research participant's concerns and preferences, which, if acknowledged by researchers and funders, will ensure that advances in research align with the values of the participants who contribute data.

This review found that although participants lacked awareness of data sharing, once given examples or vignettes, they agreed with sharing in principle.[27,28,31] They suggested that the consent process should be a tool that explains sharing[31] so that consent is not just 'informed' by name, but in practice. Although more evidence is needed to determine whether there is a causal relationship between information provision and acceptance of data sharing, it is possible that strategies promoting the translational benefits of sharing to existing or potential participants could allay fears and encourage participation (or reduce opt-out). Further research is required to determine how and by whom this promotion should be delivered, and it could be tied to the varying degrees of participant trust in different stakeholders.

Consistent with other research,[11,38] we find that some participants expressed a preference for re-consenting before sharing, and at the very least, the majority of participants preferred a thorough initial consent process with agreed terms as opposed to '*inferred consent*',[39] or inferred opt-in where researchers neglected to ask for explicit consent to share. In practice, researchers may not be able to re-contact and re-consent before secondary data use, and this review identified that participants understand these challenges.[28] However, simply giving the option to opt-out at any stage could help persuade research participants to consent to data sharing.[38] The benefits of this must be weighed against the effects of opt-out on the primary study dataset. Excluding opt-out participants means the primary data set differs to that shared, limiting reproducibility and meta-analyses.[3] Excluding participants who decline to share their data will require time on part of the researcher (in addition to the resources already required) to prepare the dataset for secondary usage, which risks missing targets for timely sharing, such those initially proposed by the ICMJE.[2]

To refine the consent process researches should seek input from patient focus groups or members of the public combined with the evidence gathered in this review.[24–27,29–31] An '*active communication*' at time of consent regarding likely future data uses[39] and types of researchers or archives with which data may be shared could be included as standard on consent forms.[26,29,30]

The review found that participants wanted adequate protection for their data and suggested that RDRs should be appropriately managed, with requests for data dealt with by committees[29] which include lay members, echoing a recommendation by the Institute of Medicine.[3] However, there are currently no hard rules or restrictions[2] regarding the location of shared data, ranging from supplementary material in a journal, within an institution's repository or in a recognized (disciplinary) archive.[40] It also remains difficult defining what exactly an RDR is.[41] It therefore is difficult to see at this time how researchers could promise RDR

management as well as community or lay representation in them. Funders and journal editors should also take account of participants' preferences regarding location of shared data and who has access to it, and consider making exceptions.[2]

Engaging participants in research that uses their own data by keeping them informed of the types of studies using their data, and the resultant outcomes (e.g. via newsletters) could enforce researcher accountability. It would ensure that the original participant's consent is honoured and that data are used in projects that align with participants' sensibilities or at least, and perhaps more realistically, benefit the population (disease or geographic) of which the participant is a part.[25–27]

Other implications for research and practice resulting from this review concern reassurance for participants that data are thoroughly anonymized prior to sharing.

Strengths and limitations

The key strength of this review is the extensive literature search, increasing the likelihood of capturing all relevant published evidence. It is the first review known to the authors concerning secondary use of trial/study data.

The included studies were found to be of reasonable or high quality, providing support for the validity of the results. There were no protocol deviations in conducting this review.

The main limitation is the paucity of studies regarding participants' attitudes towards secondary use (sharing) of trial or study data, despite broad search criteria. It is questionable whether a review containing nine studies is truly representative of participant attitudes, although it was important to consolidate the small evidence base so that further work can be based upon its conclusions.

Five of the nine included studies[25–28,31] originated from the same research team and funder, set in low- and middle-income countries. Any similarities in findings may be due to comparative methodologies or populations. Two studies by Manhas et al. used data from one research project.[29,30]

The included studies capture a wide range of countries including high-, middle- and low-income (but not the UK or Europe), potentially enhancing generalizability of findings.

Conclusions

The available literature on participant attitudes towards sharing data from clinical trials or health interventions is scant, and this review reflects that. This study identified four themes, which can be applied to policy or practice and tested with further research. Participants were clear about their conditions for data sharing, and research should move away from a culture of vague consent, which does not permit assessment by participants of the extent to which their data will be re-used, towards one of transparency. Better information about the benefits of data sharing, alongside the desired governance, may foster a willingness to share, so increasing the availability of data for secondary use.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

1. Walport M and Brest P. Sharing research data to improve public health. *Lancet* 2011; 377: 537–539.
2. Taichman DB, et al. Sharing clinical trial data: a proposal from the international committee of medical journal editors. *JAMA* 2016; 315: 467–468.
3. Institute of Medicine. *Sharing clinical trial data maximising benefits minimising risk*. Washington DC: The National Academies Press, 2015.
4. Loder E and Groves T. The BMJ requires data sharing on request for all trials. *Br Med J* 2015; 350: h2373.
5. PLOS. Data availability, http://journals.plos.org/plosone/s/data-availability (2014, accessed 14 December 2016).
6. UK Data Archive. The UK's largest collection of digital research data in the social sciences and humanities, www.data-archive.ac.uk (2015, accessed 18 March 2015).
7. Rowhani-Farid A and Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. *BMJ Open* 2016; 6: 8.
8. UK Data Service. Manage data-plan to share, https://ukdataservice.ac.uk/deposit-data/how-to (2016, accessed 07 December 2016).
9. Hill E, et al. Let's get the best quality research we can': public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. *BMC Med Res Methodol* 2013; 13: 72.
10. Stone MA, et al. Sharing patient data: competing demands of privacy, trust and research in primary care. *Br J Gen Pract* 2005; 55: 783–789.
11. Chan TW, Mackey S and Hegney DG. Patients' experiences on donation of their residual biological samples and the impact of these experiences on the type of consent given for the future research use of the tissue: a systematic review. *Int J Evid Base Health* 2012; 10: 9–26.

12. Kass NE, et al. The use of medical records in research: what do patients want? *J Law Med Ethics* 2003; 31: 429–433.
13. Nair K, et al. Patients' consent preferences regarding the use of their health information for research purposes: a qualitative study. *J Health Serv Res Policy* 2004; 9: 22–27.
14. Ludman EJ, et al. Glad you asked: participants' opinions of re-consent for dbGaP Data Submission. *J Empir Res Hum Res Ethics* 2010; 5: 9–16.
15. Robling M, et al. Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study. *J Med Ethics* 2004; 30: 104–109.
16. Lemke AA, et al. Public and biobank participant attitudes toward genetic research participation and data sharing. *Publ Health Genom* 2010; 13: 368–377.
17. Grande D, et al. Public preferences about secondary uses of electronic health information. *JAMA Intern Med* 2013; 173: 1798–1806.
18. King T, Brankovic L and Gillard P. Perspectives of Australian adults about protecting the privacy of their health information in statistical databases. *Int J Med Inform* 2012; 81: 279–289.
19. Trinidad SB, et al. Genomic research and wide data sharing: views of prospective participants. *Genet Med* 2010; 12: 486–495.
20. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; 4: 1.
21. PROSPERO. International prospective register of systematic reviews, https://crd.york.ac.uk/PROSPERO/#aboutpage (2017, accessed 06 December 2017).
22. Data Protection Commissioner. EP 95/46/EC – the data protection directive, www.dataprotection.ie/docs/EU-Directive-95-46-EC/89.htm (1995, accessed 7 December 2016).
23. CASP. Critical Appraisal Skills Framework Qualitative Appraisal Tool. CASP checklists, www.casp-uk.net/casp-tools-checklists/c18f8 (2013, accessed 01 April 2015).
24. Asai A, et al. Attitudes of the Japanese public and doctors towards use of archived information and samples without informed consent: preliminary findings based on focus group interviews. *BMC Med Ethics* 2002; 3: 10.
25. Cheah P, et al. Perceived benefits, harms, and views about how to share data responsibly: a qualitative study of experiences with and attitudes toward data sharing among research staff and community representatives in Thailand. *J Empir Res Hum Res Ethics* 2015; 10: 278–289.
26. Hate K, et al. Sweat, skepticism, and uncharted territory: a qualitative study of opinions on data sharing among public health researchers and research participants in Mumbai, India. *J Empir Res Hum Res Ethics* 2015; 10: 239–250.
27. Jao I, et al. Involving research stakeholders in developing policy on sharing public health research data in Kenya: views on fair process for informed consent, access oversight, and community engagement. *J Empir Res Hum Res Ethics* 2015; 10: 264–277.
28. Jao I, et al. Research stakeholders' views on benefits and challenges for public health research data sharing in Kenya: the importance of trust and social relations. *PLoS One* 2015; 10: e0135545.
29. Manhas KP, et al. Parent perspectives on privacy and governance for a pediatric repository of non-biological, research data. *J Empir Res Hum Res Ethics* 2015; 10: 88–99.
30. Manhas KP, et al. Parental perspectives on consent for participation in large-scale, non-biological data repositories. *Life Sci Soc Policy* 2016; 12: 1.
31. Merson L, et al. Trust, respect, and reciprocity: informing culturally appropriate data-sharing practice in Vietnam. *J Empir Res Hum Res Ethics* 2015; 10: 251–263.
32. BestBETs. BETs CA worksheets, http://bestbets.org/home/bets-introduction.php (2012, accessed 25 October 2016).
33. Platt J and Kardia S. Public trust in health information sharing: implications for biobanking and electronic health record systems. *JPM* 2015; 5: 3.
34. Thomas J and Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol* 2008; 8: 10.
35. Shabani M, Bezuidenhout L and Borry P. Attitudes of research participants and the general public towards genomic data sharing: a systematic literature review. *Expert Rev Mol Diagn* 2014; 14: 1053–1065.
36. da Silva ME, et al. Informed consent for record linkage: a systematic review. *J Med Ethics* 2012; 38: 639–642.
37. Aitken M, et al. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Med Ethics* 2016; 17: 73.
38. Taylor MJ and Taylor N. Health research access to personal confidential data in England and Wales: assessing any gap in public attitude between preferable and acceptable models of consent. *Life Sci Soc Policy* 2014; 10: 15.
39. UK Data Service. Consent for data sharing, http://data sharing.net (2012, accessed 24 January 2014).
40. Whyte A. 'Where to keep research data: DCC checklist for evaluating data repositories' *v.1.1*, www.dcc.ac.uk/resources/how-guides (2015).
41. Pampel H, et al. Making research data repositories visible: the re3data.orgRegistry. *PLoS One* 2013; 8: e78080.

Appendix B. Systematic review search terms by database

ASSIA		
Participants	**Attitudes**	**Data sharing**
(("patients" OR "patient preferences" OR "patient/patients" OR "public knowledge" OR "public awareness" OR "public acceptance" OR "public" OR "participants")	AND ("consent" OR "attitudes--beliefs" OR "privacy" OR "public opinion" OR "informed consent" OR "attitudes" OR "attitude/attitudes/attitudinal" OR "confidentiality" OR "public concerns" OR "opinion" OR "opinions" OR "public knowledge" OR "trust")	AND ("data sets" OR "datasets" OR "data" OR "data banks" OR "biodata" OR "data sources" OR "information sharing" OR "data protection" OR "patient information" OR "databases" OR "data management systems" OR "information dissemination" OR "health records" OR "medical records" OR "computerized medical records" OR "privacy" OR "information sharing" OR "information exchange" OR "health information" OR "access to information")) AND stype.exact("Scholarly Journals" OR "Trade Journals") AND pd(>19941231)) AND stype.exact("Scholarly Journals" OR "Trade Journals")

EBSCO & CINAHL		
Participants	**Attitudes**	**Data sharing**
((“patients” OR “public” OR “Research	AND (“attitude” OR “opinion” OR “Privacy” OR “confidentiality” OR “Informed consent” OR “Public	AND (“Data Sharing” OR “Clinical trials” OR “Trial data” OR “Patient data” OR “Information sharing” OR “Clinical trials” OR “Access to

subjects" OR "Patients")	opinion" OR "Privacy and confidentiality" OR "Patient attitudes" OR "Consent (research)")	information" OR "Medical records" OR "Health information" OR "Health information networks" OR "Computerized Patient Record"))

Embase		
Participants	**Attitudes**	**Data sharing**
((*Patient?" OR "*patients" OR "Research subject?" OR "*research subjects" OR "Public")	AND ("Opinion?" OR "*Public opinion" OR "understanding" OR "View?" OR "Attitude to health" OR "*attitude to health" OR "Bioethical issues" OR "*Bioethical issues")	AND ("Trial data" OR "Health information exchange" OR "*health information exchange" OR "Shar? Data"))

HMIC		
Participants	**Attitudes**	**Data sharing**
(("Patient?" OR "*patients" OR "public" OR "Patient data" OR "Patient information" OR "*patient information")	AND ("View?" OR "Informed consent" OR "*informed consent" OR "Public opinion" OR "*Public opinion" OR "Patient privacy" OR "*Patient privacy" OR "Trust" OR "Confidentiality" OR "Patient views" OR "*Patient views")	AND ("Trial data" OR "Access to information" OR "*access to information (ethics)" OR "Confidentiality" OR "*confidentiality (ethics)" OR "Electronic patient records" OR "*Electronic patient records" OR "Information exchange" OR "*information exchange"))

MEDLINE		
Participants	**Attitudes**	**Data sharing**

((“participant?” OR “patient?” OR “Patients” OR “research subject?” OR “Research Subjects” OR “public”)	AND (“Attitude” OR “attitude?” OR “Public Opinion” OR “opinion?” OR “understanding” OR “view?” OR “attitude to health” OR “Attitude to Health” OR “bioethical issues” OR “Bioethical Issues”)	AND (“data sharing” OR “trial data” OR “access to information” OR “confidentiality” OR “medical records systems” OR “health information exchange” OR “Health Information Exchange” OR “shar? Data”)).

PsychINFO		
Participants	**Attitudes**	**Data sharing**
((“participant” OR “Patients” OR “*patients” OR “Research subject?” OR “*experimental subject” OR “public”)	AND (“Attitudes” OR “attitude” OR “public opinion” OR “*Public Opinion” OR “understanding” OR “View?” OR “Informed consent” OR “patient view” OR “health attitudes” OR “Health Attitudes”)	AND (“Trial data” OR “access to information” OR “data sharing” OR “confidentiality” OR “health information exchange”))

Web of knowledge		
Participants	**Attitudes**	**Data sharing**
((“public*” OR “patient*” OR “participant*”)	AND (“view*” OR “opinion*” OR “attitude” OR “confidentiality” OR “privacy” OR “trust” OR “informed consent”)	AND (“information exchange” OR “health record” OR “patient record” OR “medical record” OR health information” OR “trial data” OR “clinical trial” OR “data sharing”))

Appendix C. List of ineligible grey literature

The table below shows ineligible grey literature identified during searches.

Author	Year	Title	Reason for ineligibility
UKCRC Registered Clinical Trials Units	2021	Considerations for a Participant Data Sharing SOP	Guidance on preparation of a SOP not guidance on sharing as such
Pierce et al (Association of American Medical Colleges)	2019	Supplementary information to: Credit data generators for data reuse (To accompany a Comment published in Nature 570, 30–32)	Non-UK American
Office for statistics regulation	2019	NHS Digital data sharing and access review: initial conclusions and scope for further review	Concerns healthcare data
Association of American Medical Colleges (Heather H. Pierce, Anurupa Dev, Emily Statham and Barbara E. Bierer)	2018	Implementing a System to Enable Credit for Data Sharing (Supplementary information to: Credit data generators for data reuse To accompany a Comment published in Nature)	Non-UK American
Steve Olson and Autumn S. Downey	2017	Sharing Clinical Research Data- workshop summary	Non-UK American
Academy of Medical Sciences	2017	Personal data for public good: using health information in medical research	No guidance given (data sharing mentioned only once)

Modjarrad K, Moorthy VS, Millett P, Gsell PS, Roth C, Kieny M-P (World Health Organization)	2016	Developing Global Norms for Sharing Data and Results during Public Health Emergencies	Collaboration with countries outside of UK
Research Data Alliance	2015	Outputs	Collaboration with countries outside of UK
The National Academies	2015	Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk	Non-UK American
Applied Clinical Trials and Pharmaceutical Executive	2014	Clinical trial data sharing and disclosure	Non-UK American
National Institute of Health	2014	Genomic Data Sharing Policy	Non-UK American
European Research Council	2014	Open Access Guidelines for research results funded by the ERC	Non-UK European
European Research Council	2014	Workshop on Research Data Management and Sharing- Abstracts of the presentations	Non-UK European
The World Medical Association, inc.	2013	Proposed WMA Declaration on Ethical Considerations regarding Health Databases and Biobanks	Collaboration with countries outside of UK
G8UK	2013	G8 Science Ministers Statement London UK	Collaboration with countries outside of UK
Dame Fiona Caldicott (gov.uk)	2013	Information: To share or not to share? The Information Governance Review	Too high level
Association of Medical Research Charities	2013	Statement on the use of patient data for research	Concerns healthcare data
Biotechnology and Biological Sciences Research Council (BBSRC)	2010	BBSRC data sharing policy Frequently asked Questions	Relevant to Biosciences

Biotechnology and Biological Sciences Research Council (BBSRC)	2010	Data Sharing in the Biosciences	Relevant to Biosciences
Biotechnology and Biological Sciences Research Council (BBSRC)	2010	BBSRC data sharing policy	Relevant to Biosciences
Academy of Medical Sciences	2008	Submission to Data Sharing Review	No guidance given
Organisation for economic co-operation and development	2007	OECD Principles and Guidelines for Access to Research Data from Public Funding	Collaboration with countries outside of UK

Appendix D. Scoping focus group topic guide

Remind participants that everything they say will be confidential and anonymised. Provide disclaimer that in the unlikely event that anything is shared that could put themselves or others at risk, I will be bound to share this information.

Background

Introductions- give participants info about me and research and purpose of research.

What do I mean by data? (public health research/trial data).

What do I mean by data sharing? - data sharing policy increasing now, secondary use of data.

Have participants been involved in any trials/interventions/research? What kind?

Awareness of data sharing

Have participants heard of data sharing?

Did participants know what it meant?

Have participants ever consented to data sharing/been involved in research where data could be shared?

General views on data sharing

Can you see any advantages of data sharing? What might those advantages be?

Who might benefit from data sharing? How? Why? Would data sharing benefit the participants themselves?

Can you identify any disadvantages of data sharing? What might those disadvantages be?

Who might be adversely affected? How? Why? How might participants themselves be adversely affected?

Are there any types of data that could be of particular benefit to share? Why?

Any types of data that we should be particularly careful about sharing/shouldn't share at all? Why? (sensitive data) what is sensitive?

Experiences of data sharing (if applicable)

If applicable, can you describe their experience of data sharing from consent to study end.

How do you feel about your data being shared and re-used?

Consent preferences

Explain that consent must be (should be) given to share participants' data. Use earlier example if required

Do you think consent is necessary? Are there any circumstances in which data could be shared without consent? If so, what circumstances?

How would you feel if your data were shared without permission?

Would you prefer that consent for sharing was given once at the beginning of the original study (broad consent), or on a case-by-case basis (i.e., for each subsequent study that wanted to use the data)?

What advantages and disadvantages of each approach can you see?

Do you have any thoughts about how permission for data sharing could be incorporated into consent process?

Governance issues/preferences

How do you think that data should be stored ready for sharing?

Think about access, anonymisation, processing requests for data, who it can be shared with, researcher responsibilities, participant interests, any constraints.

Sharing data from this focus group/interview?

Explain how the data will be used- for development of questionnaires and summarized and reported in a thesis, and potentially a paper to be published.

Is participant still happy for data given in these focus groups/ interviews to be used? To be shared?

For each focus group/interview record the date, start and finish times, gender and age of participants and any comments/reflections.

Appendix E. Some examples of how Cognitive Interviews affected Questionnaire wording and structure

Question/section	Text before cognitive interviewing	Interviewee Comment	Text after cognitive interviewing	Reasoning/ outcome
Survey background- *"human participants"*	*"A clinical trial is a medical study in which* ***human*** *participants receive treatment according to a research plan."*	Why the word human? Word human not necessary	*"A clinical trial is a medical study in which participants receive treatment according to a research plan."*	Unnecessary word removed.
Survey background- *"these treatments"*	*"These treatments could be drugs; medical procedures;* ***or changes to participants' behaviour****, such as their diet."*	I wouldn't put diet down as a behaviour. Put behaviour first.	*"These treatments could be medicines; medical procedures; or* ***attempts to change participant's' behaviour****, such as their diet."*	Added in *"attempts to change"* behaviour such as diet.
Survey background- data sharing	*"****Data sharing means removing personal identifiers*** *(like names and birthdates) from the information that is collected about participants and then allowing other researchers (who aren't part of the original research team) to*	Data sharing doesn't mean removing personal identifiers. It means sharing.	*"****Data sharing means allowing other researchers (who aren't part of the original research team) to see and use study data for further research. Personal identifiers (like names and***	Rearranged sentence to explain data sharing and then that identifiers should be removed.

	see and use the data for further research."		***birthdates) should be removed** so that the data is anonymous."*	
Q1	*"Have any of the following ever taken part**/are currently taking part** in a health research study?"*	Clunky phrasing. Consider or instead of /	*"Have any of the following ever taken part in a health research study?"*	Deleted *"are currently taking part"* as question asks *"if ever"*, which would include now.
Questions about data sharing- instructions for completion	*"IF YOUR CHILD **OR SOMEONE ELSE close to you** has taken part in a health research study, please answer the following questions thinking about sharing their data."*	"If your child or someone else you know" - I think this is incorrect – you are asking for third- or fourth-hand information – it can't be accurate. I think you should only ask for "your child" or "if your child or a close family member"	*"**IF YOUR CHILD** has taken part in a health research study, please answer the following questions thinking about sharing their data."*	Changed questionnaire to ask only about *"you"* or *"your child"* for simplicity and accuracy.
Q7	*"My data could be used in research I don't approve of"*	Should the word 'if' be used in the options?	*"**If** my data could be used in research I don't approve of"*	Added 'if' to most options as it sounds more colloquial.
Q8	*"To help government study health problems"*	Which government? Ours or EU?	*"To help **the** government study health problems"*	Changed to *"the government"*

				implying that it is the UK government.
Q9-c	***"It can help patients by getting*** *quicker answers to scientific questions using data already collected."*	This sentence is confusing	***"Researchers can get quicker answers to scientific questions*** *using data already collected."*	Changed text. Hopefully it is obvious this would then benefit patients.
Q9-e	*"I can contribute* ***more data*** *to research that affects me or my family."*	How?	*"I can contribute* ***to more*** *research that affects me or my family."*	Re-worded
Q10	n/a	Nothing about civic responsibility. For scientific knowledge. Or to be nice.	*"Chance to help others by contributing to research"*	Added option 10e.
Q11-	*"How willing would you be for them to share anonymised details of:"*	Do we need a reminder that this would be anonymous sharing? A while since it was mentioned.	*"How willing would you be for them to share* ***anonymised*** *details of:"*	Added word *"anonymised"* to q 11 itself. Anonymised is explained in other places in the questionnaire.
Q19-22	*19 How would you prefer your study data to be stored?*	…presumes an intuitive understanding of issues to do with storage etc (e.g., Q19-Q22)'	*19 How would you prefer your study data to be stored?*	Amended this section to make it shorter and simpler.

Q19 –Q23	***20 Imagine the data was stored with 'controlled access' (there is a formal request and approval process in place).*** ***Where would you prefer your data to be stored prior to it being shared?*** ***21 Do you think there should be a limit to the number of times study data should be shared?*** *22 Who do you think should 'own' **(control access to)** the data collected during a study?* *23 Do you think it is important that researchers using people's shared data give feedback telling participants how their data was used?*	Too complicated for man in the street.	***20 If data has controlled access:*** ***Who do you think should give permission for data to be shared and used again?*** *21 Who do you think should 'own' the data collected during a study?* *22 Do you think it is important that researchers using shared data give feedback telling participants how their data was used?*	
Q20- *"with the original researcher who collected it"*	***"Imagine the data was stored with 'controlled access' (there is a***	Might be a group or institution. Database run by university could still be with original researcher.	***"If data has controlled access:***	Question removed. Simpler replacement

	formal request and approval process in place). ***Where would you prefer your data to be stored prior to it being shared?"*** ☐ With the **original** Researcher who collected it ☐ **In an online database run by a University** ☐ **In an online database run by a specialist organisation** ☐ **In an offline database** ☐ **I'm not comfortable with it being stored anywhere** ☐ Other ☐ Not Sure	Clarify as some of these options could be some of the same thing	***Who do you think should give permission for data to be shared and used again?"*** ☐ **The participants who took part should decide** ☐ **The researcher(s) who collected it** ☐ **The organisation where the original researcher(s) work** ☐ **An independent committee** ☐ Other ☐ Not sure	question and options added.
Q22	*"Who do you think should 'own' **(control access to)** the data collected during a study?"*	Owning and controlling access to is not the same thing/person	*"Who do you think should 'own' the data collected during a study? "*	Removed *"control access to."*

Appendix F. Some examples of how readability testing affected Questionnaire wording and structure

The table below shows how readability testing on a questionnaire post cognitive interviewing prompted changes to questionnaire text.

<table>
<tr><th>Section text</th><th>Readability consensus Scott 2017 readabilityformulas.com</th><th>New readability Consensus Scott 2017 readabilityformulas.com</th></tr>
<tr><td>About this survey</td><td>Grade Level: 11

Reading Level: fairly difficult to read.

Reader's Age: 15-17 yrs. old (Tenth to Eleventh graders)</td><td>Grade Level: 11

Reading Level: fairly difficult to read.

Reader's Age: 15-17 yrs. old (Tenth to Eleventh graders)</td></tr>
<tr><td colspan="3">Text before readability testing:

About this Survey:
• We are interested in how members of the public or people who have taken part in research feel about what researchers do with their data at the end of the study.
• When participants take part in a research study, they are asked to sign a consent form. The consent form confirms that the participant is happy to include their data in that research study, and may ask if it is ok to share the data with other researchers at the end of the study.
We would like to understand how participants feel about researchers sharing their data with other researchers.
It is hoped that by understanding participants preferences for data sharing, researchers could change the way in participants are told about data sharing, or how they are asked to share data.</td></tr>
<tr><td colspan="3">Text after readability testing:</td></tr>
</table>

About this Survey:

- We are interested in how members of the public, or people who have taken part in research feel about what researchers do with their data at the end of the study.
- When participants take part in a research study, they are asked to sign a consent form. The consent form checks that the participant is happy to include their data in that research study, and may ask if it is ok to share the data with other researchers at the end of the study.
- We would like to understand how participants feel about researchers sharing their data with other researchers.
- We hope that by understanding participants preferences for data sharing, researchers could change the way participants are told about data sharing, or how they are asked to share data.

Section text	**Readability consensus Scott 2017 readabilityformulas.com**	**New readability Consensus Scott 2017 readabilityformulas.com**
Survey background	Grade Level: 11 Reading Level: difficult to read. Reader's Age: 15-17 yrs. old (Tenth to Eleventh graders)	Grade Level: 12 Reading Level: difficult to read. Reader's Age: 17-18 yrs. old (Twelfth graders)
Text before readability testing:		

Survey Background

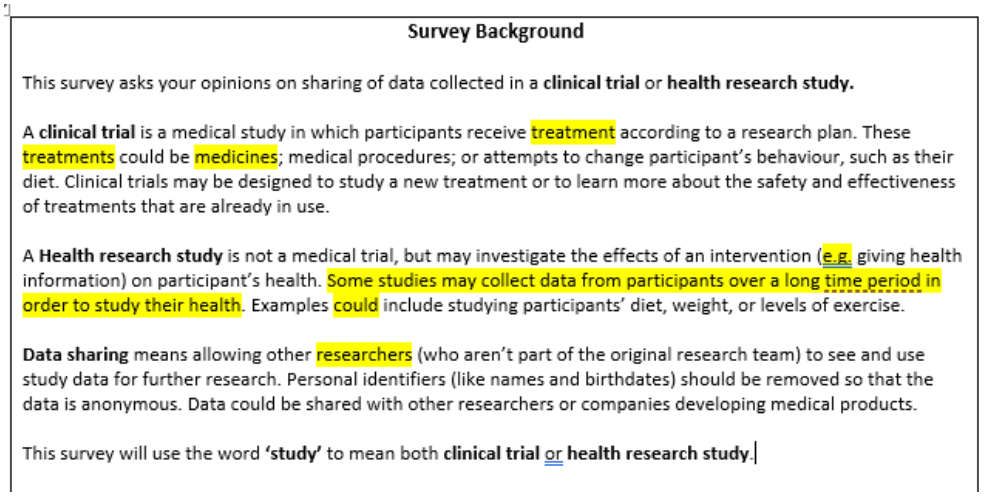
This survey asks your opinions on sharing of data collected in a **clinical trial** or **health research study.**

A **clinical trial** is a medical study in which human participants receive specific interventions according to a research plan. These interventions could be medical products such as drugs; medical procedures; or changes to participants' behaviour, such as diet. Clinical trials may be designed to study a new intervention or to learn more about the safety and effectiveness of interventions that are already in use.

A **Health research study** is not a medical trial but may investigate the effects of an intervention on health (such as giving participants information) or may collect data from participants over a long time period in order to draw conclusions. Examples include studying participants diet, weight, or levels of exercise.

Data sharing means removing personal identifiers (like names and birthdates) from the information that is collected about participants and then allowing other people who aren't part of the original research team to see and use the data for further research. Data could be shared with other researchers or companies developing medical products. This survey will use the word 'study' to mean both research study or clinical trial.

Text after readability testing:



Survey Background

This survey asks your opinions on sharing of data collected in a **clinical trial** or **health research study.**

A **clinical trial** is a medical study in which participants receive treatment according to a research plan. These treatments could be medicines; medical procedures; or attempts to change participant's behaviour, such as their diet. Clinical trials may be designed to study a new treatment or to learn more about the safety and effectiveness of treatments that are already in use.

A **Health research study** is not a medical trial, but may investigate the effects of an intervention (e.g. giving health information) on participant's health. Some studies may collect data from participants over a long time period in order to study their health. Examples could include studying participants' diet, weight, or levels of exercise.

Data sharing means allowing other researchers (who aren't part of the original research team) to see and use study data for further research. Personal identifiers (like names and birthdates) should be removed so that the data is anonymous. Data could be shared with other researchers or companies developing medical products.

This survey will use the word **'study'** to mean both **clinical trial** or **health research study**.

Your views on sharing data from clinical trials and health research studies

Start of Block: Patient Information Sheet

Your views on sharing data from clinical trials and health research studies

About this survey:

We are interested in how members of the public, or people who have taken part in research, feel about what researchers do with their data at the end of the study. When participants take part in a research study, they are asked to sign a consent form. The consent form checks that the participant is happy to have their data included in the research study. Sometimes the consent form asks permission to share the data with other researchers at the end of the study. We would like to understand how participants feel about researchers sharing their data with other researchers. We hope that by understanding participant's preferences for data sharing, researchers could change the way participants are told about data sharing, or how they are asked to consent to share data.

What does taking part involve?

Taking part in the study involves completing the attached survey. It should take approximately 20 minutes to complete this survey. You can also save your answers and continue the survey at another time.

Your survey responses will be given a study number and saved in a database. The survey data will be analysed with the responses from the other people completing the survey. There are no wrong or right answers - we are just interested in your opinion and any relevant experiences.

Risks and Benefits:

Participating in this study has no direct benefits to you, although you may find it rewarding to contribute to research. We will take steps to ensure that your survey responses remain confidential. Your data will only be identified by a study number and your data will be stored securely. This study was approved by the Faculty of Medical Sciences Research Ethics Committee, part of Newcastle University's Research Ethics Committee. This committee contains members who are internal to the Faculty, as well as one external member. This study was reviewed by members of the committee, who must provide impartial advice and avoid significant conflicts of interests.

Consent:

By completing this survey, you are giving your informed consent to take part. Completing this survey will not affect the treatment you will receive in any study or trial in which you are already involved. It won't affect what happens to data from any studies or trials you may have been involved in. The results of this survey may be presented at scientific meetings or published in scientific journals. The de-identified survey data may be shared with other researchers if it is requested, and if approval is granted. The survey data may be shared with the original study team who contacted you to take part (if applicable).

For more information, or if you have any concerns about the study please contact:

Nicola Howe (lead researcher) Newcastle Clinical Trials Unit, 1-4 Claremont Terrace, Newcastle University, Newcastle upon Tyne, NE2 4AE, UK.
Nicola.howe@newcastle.ac.uk Tel: 0191 208 8024 **OR** Professor Elaine McColl, Professor of Health Service Research (lead research supervisor). Institute of Health and Society, Newcastle University, Baddiley-Clark Building, Richardson Road, Newcastle upon Tyne, NE2 4AX. Elaine.McColl@newcastle.ac.uk Tel: 0191 208 7260

End of Block: Patient Information Sheet

Start of Block: Questions about taking part in research

Survey Background

This survey asks your opinions on sharing of data collected in a clinical trial or health research study.

A clinical trial is a medical study in which participants receive treatment according to a research plan. These treatments could be medicines; medical procedures; or attempts to change participant's' behaviour, such as their diet. Clinical trials may be designed to study a new treatment or to learn more about the safety and effectiveness of treatments that are already in use.

A Health research study is not a medical trial but may investigate the effects of an intervention (e.g., giving health information) on participant's health. Some studies may collect data from participants over a long time period in order to study their health. Examples could include studying participants' diet, weight, or levels of exercise.

Data sharing means allowing other researchers (who aren't part of the original research team) to see and use study data for further research. Personal identifiers (like names and birthdates) should be removed so that the data is anonymous. Data could be shared with other researchers or companies developing medical products. This survey will use the word 'study' to mean both clinical trial or **health re**search study.

Q1

Questions about taking part in research

Have any of the following ever taken part in a health research study?

(select all that apply)

- ☐ You (1)
- ☐ Your child (2)
- ☐ ⊗ Neither (3)
- ☐ ⊗ Not sure (4)

Display This Question:

If Q1 = You

Q2a

Was YOUR participation in the study as:

(please select one)

- ○ A person who had the health condition being studied? (1)
- ○ A healthy volunteer? (2)
- ○ A person who is at risk of developing the health condition being studied? (3)
- ○ Not sure (4)

Display This Question:

If Q1 = Your child

Q2b **Was YOUR CHILD'S participation in the study as:**

(please select one)

- ○ A child who had the health condition being studied? (1)
- ○ A healthy volunteer? (2)
- ○ A child who is at risk of developing the health condition being studied? (3)
- ○ Not sure (4)

Display This Question:

If Q1 = You

Q3a

What was the experience of taking part in a study like for YOU?

(please select one)

- ○ Very positive (1)
- ○ Positive (2)
- ○ Neither positive or negative (3)
- ○ Negative (4)
- ○ Very negative (5)
- ○ Not applicable (6)
- ○ Not sure (7)

Display This Question:

If Q1 = Your child

Q3b **What was the experience of taking part in a study like for YOUR CHILD?**

(please select one)

- ○ Very positive (1)
- ○ Positive (2)
- ○ Neither positive nor negative (3)
- ○ Negative (4)
- ○ Very negative (5)
- ○ Not applicable (6)
- ○ Not sure (7)

End of Block: Questions about taking part in research

Start of Block: Questions about Data Sharing

Display This Question:

If Q1 = Not sure

Or Q1 = Neither

Instructions for completion

If you have NOT taken part in a health research study, please answer the questions as if you had taken part or are going to take part- think about your study data being shared.

Page Break

End of Block: Questions about Data Sharing

Start of Block: Questoins 5 and 6

Questions about Data Sharing

Remember, sharing study data means sharing information about each individual research participant- for example, age, health conditions, and response to the treatment being tested-not just the overall results of the study.
The data will be anonymised- personal identifiers (like names and birth dates) would be removed. Data could be shared with other researchers or with companies developing medical products.

Q5 **How concerned would you be if you knew data from a study that you were involved in was being shared?**

(please select one)

- ○ Very concerned (1)
- ○ Somewhat concerned (2)
- ○ Not very concerned (3)
- ○ Not at all concerned (4)
- ○ Not sure (5)
- ○ Depends who it is shared with (6)

Q6 How concerned would you be if you knew data was being shared with:

	Very concerned (1)	Somewhat concerned (2)	Not very concerned (3)	Not at all concerned (4)	Not sure (5)
Researchers at the same organisation where your data was collected (1)	○	○	○	○	○
Researchers at a pharmaceutical company, e.g. for developing new medicines (2)	○	○	○	○	○
Researchers at a university (3)	○	○	○	○	○
Researchers at a hospital (4)	○	○	○	○	○
Researchers in another country (5)	○	○	○	○	○
A charity or not for profit organisation (6)	○	○	○	○	○
The government (7)	○	○	○	○	○

A student at a university (8)	○	○	○	○	○
On the internet for anyone to use (9)	○	○	○	○	○

End of Block: Questoins 5 and 6

Start of Block: Q7

Q7 **If data from a study in which you were involved was being shared, how concerned would you be about the following?**

	Very concerned (1)	Somewhat concerned (2)	Not very concerned (3)	Not at all concerned (4)	Not sure (5)
a) If I could still be identified in the data (1)	○	○	○	○	○
b) If my data could be used in research I don't approve of (2)	○	○	○	○	○
c) If my data could be stolen (3)	○	○	○	○	○
d) If my data could be used for making a profit e.g., advertising instead of research (4)	○	○	○	○	○
e) If it would be embarrassing if my data was linked back to me (5)	○	○	○	○	○

f) If people could misinterpret the data and come to the wrong conclusions (6)	○	○	○	○	○
g) If the original research team didn't get credit for collecting the data (7)	○	○	○	○	○
h) If it stopped researchers doing their own original research (8)	○	○	○	○	○

Q7a **Which of the above statements is of MOST concern to you?**

(please select one, a-h)

- ○ a- If I could still be identified from the data (1)
- ○ b- If my data could be used for research I don't approve of (2)
- ○ c- If my data could be stolen (3)
- ○ d- If my data could be used for making profit (4)
- ○ e- If it would be embarrassing if my data was linked back to me (5)
- ○ f- If people could misinterpret the data (6)
- ○ g- If the original research team didn't get credit (7)
- ○ h- If it stopped researchers doing original research (8)

End of Block: Q7

Start of Block: Q8

Q8 **How likely would you be to give permission for your data to be shared for the following reasons?**

	Very likely (1)	Somewhat likely (2)	Neither likely nor unlikely (3)	Somewhat unlikely (4)	Very unlikely (5)
To do research in a University (1)	○	○	○	○	○
To do research in a hospital (2)	○	○	○	○	○
To help a pharmaceutical company do research (3)	○	○	○	○	○
To help the government study health problems (4)	○	○	○	○	○
To inform the public about a health issue. (5)	○	○	○	○	○
To help students get data for projects (6)	○	○	○	○	○

End of Block: Q8

Start of Block: Questions 9 and 10

Q9 **Below is a list of potential benefits of data sharing. Which of these make you feel more positive about data sharing?** ***(please select all that apply)***

☐ Researchers can check each other's results and conclusions, making science more open (1)

☐ Rarer diseases and conditions can be studied more easily using combined data, without having to wait for more studies. (2)

☐ Researchers can get quicker answers to scientific questions using data already collected. (3)

☐ Researchers can get the most out of participant's contribution (data) to their studies. (4)

☐ I can contribute to more research that affects me or my family. (5)

Q10

Would any of the following motivate you to allow your data to be shared?

	Yes (1)	No (2)	Not sure (3)
Assured anonymity of the data shared (1)	○	○	○
Understanding exactly how the data will be used (2)	○	○	○
Knowing exactly who will access the data (3)	○	○	○
Chance to understand my own condition better (4)	○	○	○
Chance to help others by contributing to research (5)	○	○	○

End of Block: Questions 9 and 10

Start of Block: how willing to share details of...

Q11

Imagine that the researcher from the study you took part in wants to share your data with other researchers.

How willing would you be for them to share anonymised details of your:

	Not at all willing (1)	Not very willing (2)	not sure (3)	Willing (4)	Very willing (5)
Age (Q11_1)	○	○	○	○	○
Gender (Q11_2)	○	○	○	○	○
Education (Q11_3)	○	○	○	○	○
Employment (Q11_4)	○	○	○	○	○
Height & weight (Q11_5)	○	○	○	○	○
Mental health (Q11_6)	○	○	○	○	○
Cancers (Q11_7)	○	○	○	○	○
HIV infection (Q11_8)	○	○	○	○	○
Other diseases or conditions (Q11_9)	○	○	○	○	○
Family history of disease (Q11_10)	○	○	○	○	○

Reproductive health (Q11_11)	○	○	○	○	○
Medications being taken (Q11_12)	○	○	○	○	○
Smoking behaviour (Q11_13)	○	○	○	○	○
Alcohol use (Q11_14)	○	○	○	○	○
Illegal drug use (Q11_15)	○	○	○	○	○

End of Block: how willing to share details of...

Start of Block: Questions about consent

Questions about Consent Researchers should let participants know on the consent form that their study data could be shared. The consent form is signed by participants (or parents/guardians of participants) at the beginning of a study.

Q12 **How and when would you like to be asked to share your data?**

(please select one)

- ○ Once, on the consent form for the original study (1)
- ○ Every time it is shared, with the option for me to say no (2)
- ○ Just let me know every time it is shared (3)
- ○ There is no need to ask me, just share it (4)
- ○ I have no preference (5)

Q13

What information would you like to see on the consent form before you agree to share your data?

(please tick yes, no or not sure for each statement)

	Yes (1)	No (2)	Not sure (3)
Explain that my data may be shared (1)	○	○	○
HOW the researchers will protect (anonymise) my identity. (8)	○	○	○
Explanation of WHO might benefit from using my data (3)	○	○	○
Details of WHERE the data will be stored. (4)	○	○	○
Details of HOW the data will be stored. (5)	○	○	○
Details of WHO the data might be shared with. (6)	○	○	○

Q13a **OR**

○ None of the above would convince me to share my data (1)

End of Block: Questions about consent

Start of Block: questions about consent 2

Q14 **How important is it that you are informed on the consent form that your study data might be shared?**

(please select one)

- ○ Very important (1)
- ○ Somewhat important (2)
- ○ Not very important (3)
- ○ Not at all important (4)
- ○ Not sure (5)

Q15 **If you knew your data might be shared, what effect would it have on you taking part in a study?**

(please select one)

- ○ I would not take part at all (1)
- ○ I'd be much less likely to take part (2)
- ○ I'd be a bit more cautious about taking part (3)
- ○ It would have no effect on my decision to take part (4)
- ○ I'd be a bit more likely to take part (5)
- ○ I'd be much more likely to take part (6)
- ○ Not sure (7)

Q16 **Would you prefer to give consent separately for each type of organisation your data could be shared with?** ***(For example, separate consent for within the NHS, within the university doing the original research, outside the university, private companies).***

- ○ Yes (1)
- ○ No (2)
- ○ Not sure (3)

End of Block: questions about consent 2

Start of Block: Questions about register

Q17 **Do you think a register of participants willing to share their study data is a good idea?**

(researchers could refer to the register instead of having to gain consent for each study the participant takes part in)

- ○ Yes (1)
- ○ No (2)
- ○ Not sure (3)

Q18 **If a register of participants who are willing to share their study data existed, would you be willing to be named on it?**

- ○ Yes (1)
- ○ No (2)
- ○ Not sure (3)

End of Block: Questions about register

Start of Block: Questions about data storage

Questions about data storage **Data can be stored with *controlled access* or *open access*.**

With **controlled access,** there is a formal request and approval process in place before data can be shared. With **open access,** data can be accessed by anyone.

Q19 **How would you prefer your study data to be stored?**

(please select one)

- ○ Open access (1)
- ○ Controlled access (2)
- ○ No preference (3)
- ○ Not sure (4)

Q20 **If data has controlled access: Who do you think should give permission for data to be shared and used again?** ***(please select one)***

- ○ The participants who took part should decide (1)
- ○ The researcher(s) who collected it (2)
- ○ The organisation where the original researcher(s) work (3)
- ○ An independent committee (4)
- ○ Other (5)
- ○ Not sure (6)

End of Block: Questions about data storage

Start of Block: ownership and feedback

Q21 **Who do you think should 'own' the data collected during a study?** ***(please tick all that apply)***

- ☐ Me/the participants who took part (1)
- ☐ The researcher(s) who collected it (2)
- ☐ The organisation where the original researcher(s) work (3)
- ☐ Anyone who uses it (4)
- ☐ Whoever stores it (5)
- ☐ No one (6)
- ☐ Other (7)
- ☐ Not sure (8)

Q22 **Do you think it is important that researchers using shared data give feedback telling participants how their data was used?** ***(please tick one)***

- ○ Yes (1)
- ○ No (2)
- ○ Not sure (3)

End of Block: ownership and feedback

Start of Block: Questions about you

Q23

Questions about you

What is your gender?

- ○ Male (1)
- ○ Female (2)
- ○ Other (3)
- ○ Prefer not to say (4)

Q24 **Which age group do you belong to?**

- ○ 18-24 (1)
- ○ 25-44 (2)
- ○ 45-64 (3)
- ○ 65-74 (4)
- ○ 75-84 (5)
- ○ 85 and over (6)
- ○ Prefer not to say (7)

Q25 **How would you describe your ethnicity?**

- ○ White (1)
- ○ Mixed/multiple ethnic group (2)
- ○ Asian/Asian British (3)
- ○ Black/African/Caribbean/Black British (4)
- ○ Chinese (5)
- ○ Other ethnic group (6)
- ○ Prefer not to say (7)

Q26 **What is your highest level of educational achievement?** ***(please tick one)***

- ○ No qualifications (1)
- ○ O Levels/CSE/GCSE or equivalent (2)
- ○ AS/A Levels or equivalent (3)
- ○ Degree (e.g., BA, BSc) or equivalent (4)
- ○ Higher degree (e.g., MSc, PhD) or equivalent (5)
- ○ Professional qualifications (e.g., nursing, accountancy, teaching) (6)
- ○ Other (7)
- ○ Prefer not to say (8)

Q27 **How would you describe your overall health at the moment?** ***(please tick one)***

- ○ Excellent (1)
- ○ Good (2)
- ○ Average (3)
- ○ Poor (4)
- ○ Very poor (5)
- ○ Not sure (6)
- ○ Prefer not to say (7)

Q28 **What is your postcode?**

(this question is optional)

_____

Q29 **Do you have any further comments about data sharing or about this survey?**

(***if yes, please use the space below)***

_____

_____

_____

_____

_____

Q30 **Would you be interested in taking part in further research about data sharing, for example as part of a focus group or by taking part in an interview?**

○ If yes, you can contact the lead researcher by email: Nicola.howe@newcastle.ac.uk (4)

End of Block: Questions about you

Appendix H. Example invitation to take part letter

Your views on sharing data from clinical trials and health research studies

I am contacting you because you are a member of the Aberdeen Children of the 1950s study.

I am a PhD student at Newcastle University, and I would like to invite you to take part in a questionnaire survey about your views of data sharing.

We are interested in how people who have taken part in research, feel about what researchers do with their data at the end of the study.

When participants take part in a research study, they are asked to sign a consent form. The consent form checks that the participant is happy to have their data included in the research study. Sometimes the consent form asks permission to share the data with other researchers at the end of the study.

We would like to understand how participants feel about researchers sharing their data with other researchers.

We hope that by understanding participant's preferences for data sharing, researchers could change the way participants are told about data sharing, or how they are asked to consent to share data. The results of the survey will also contribute to my PhD thesis.

What does taking part involve?

Taking part in the study involves completing the survey accessed using the link provided. There are no wrong or right answers - we are just interested in your opinion and any relevant experiences.

Further details can be found in the patient information at the front of the survey, or, if you have any questions, you can contact:

Nicola Howe (lead researcher)

Newcastle Clinical Trials Unit, 1-4 Claremont Terrace, Newcastle University, Newcastle upon Tyne, NE2 4AE, UK. Nicola.howe@newcastle.ac.uk Tel: 0191 208 8024

Yours sincerely,

Nicola Howe

Appendix I. Variable Dichotomisation

The table below shows how the Likert scale responses for each variable were dichotomised.

Dependent variable	Answers	Dichotomous answers
Q5 How concerned would you be if you knew data from the study that you are involved in was being shared?	• very concerned • somewhat concerned • not very concerned • not at all concerned • not sure • depends who it is shared with	Concerned • very concerned • somewhat concerned • depends who it is shared with • not sure Not concerned • not very concerned • not at all concerned
Q6 a) Researchers at the same organisation where your data was collected b) Researchers at a pharmaceutical company, e.g., for developing new medicines c) Researchers at a university d) Researchers at a hospital e) Researchers in another country f) A charity or not for profit organisation g) The government h) A student at a university i) On the internet for anyone to use	very concerned somewhat concerned not very concerned not at all concerned not sure	Concerned • very concerned • somewhat concerned • not sure Not concerned • not very concerned • not at all concerned
Q7 If data from the study in which you were involved was being shared, how concerned would you be about the following? a) If I could still be identified in the data b) If my data could be used in research I don't approve of c) If my data could be stolen d) If my data could be used for making a profit e.g., advertising instead of research e) If it would be embarrassing if my data was linked back to me f) If people could misinterpret the data and come to the wrong conclusions	very concerned somewhat concerned not very concerned not at all concerned not sure	Concerned • very concerned • somewhat concerned • not sure Not concerned • not very concerned • not at all concerned

g) If the original research team didn't get credit for collecting the data h) If it stopped researchers doing their own original research		
Q7a Which of the above statements is of MOST concern to you?		For each statement a-h of most concern: Yes OR No
Q8 How likely would you be to give permission for your data to be shared for the following reasons? 1. To do research in a University 2. To do research in a hospital 3. To help a pharmaceutical company do research 4. To help the government study health problems 5. To inform the public about a health issue. 6. To help students get data for projects	very likely somewhat likely neither likely or unlikely somewhat unlikely very unlikely	Likely • very likely • somewhat likely Unlikely • somewhat unlikely • very unlikely • neither likely or unlikely
Q9 Below is a list of potential benefits of data sharing. Which of these make you feel more positive about data sharing?	• Researchers can check each other's results and conclusions, making science more open • Rarer diseases and conditions can be studied more easily using combined data, without having to wait for more studies. • Researchers can get quicker answers to scientific questions using data already collected. • Researchers can get the most out of participant's contribution (data) to their studies. • I can contribute to more research that affects me or my family.	For each statement: Yes OR No
Q10 Would any of the following motivate you to allow your data to be shared? 1. Assured anonymity of the data shared 2. Understanding exactly how the data will be used 3. Knowing exactly who will access the data 4. Chance to understand my own condition better	Yes No Not sure	Yes • Yes No • No • Not sure

5. Chance to help others by contributing to research		
Q11 Imagine that the researcher from the study you took part in wants to share your data with other researchers. How willing would you be for them to share anonymised details of your: 1. Age 2. Gender 3. Education 4. Employment 5. Height & weight 6. Mental health 7. Cancers 8. HIV infection 9. Other diseases or conditions 10. Family history of disease 11. Reproductive health 12. Medications being taken 13. Smoking behaviour 14. Alcohol use 15. Illegal drug use	Not at all willing Not very willing not sure Willing Very willing	Willing • Willing • Very willing Not willing • Not at all willing • Not very willing • Not sure
Q12 how and when would you like to be asked to share your data?	• Once, on the consent form for the original study • Every time it is shared, with the option for me to say no • Just let me know every time it is shared • There is no need to ask me, just share it • I have no preference	For each statement of who owns it: Yes OR No
Q13 What information would you like to see on the consent form before you agree to share your data? 1. Explain that my data may be shared 2. HOW the researchers will protect (anonymise) my identity. 3. Explanation of WHO might benefit from using my data 4. Details of WHERE the data will be stored. 5. Details of HOW the data will be stored. 6. Details of WHO the data might be shared with.	Yes No Not sure	Yes • Yes • Not sure No • No
Q13a None of the above would convince me to share my data		Yes No

Q14 How important is it that you are informed on the consent form that your study data might be shared?	Very important Somewhat important Not very important Not at all important Not sure	Important • Very important • Somewhat important • Not sure Not important • Not very important • Not at all important
Q15 If you knew your data might be shared, what effect would it have on you taking part in a study?	• I would not take part at all • I'd be much less likely to take part • I'd be a bit more cautious about taking part • It would have no effect on my decision to take part • I'd be a bit more likely to take part • I'd be much more likely to take part • Not sure	Unlikely • I would not take part at all • I'd be much less likely to take part • I'd be a bit more cautious about taking part • Not sure Likely • It would have no effect on my decision to take part • I'd be a bit more likely to take part • I'd be much more likely to take part
Q16 Would you prefer to give consent separately for each type of organisation your data could be shared with? (For example, separate consent for within the NHS, within the university doing the original research, outside the university, private companies).	Yes No Not sure	Yes • Yes No • No • Not sure
Q17 In theory, do you think a register of participants willing to share their study data is a good idea?	Yes No Not sure	Yes • Yes No • No • Not sure
Q18 In theory, if a register of participants who are willing to share their study data existed, would you be willing to be named on it?	Yes No Not sure	Yes • Yes No • No • Not sure
Q19 How would you prefer your study data to be stored?	Open access Controlled access No preference Not sure	Open access • Open access • No preference Controlled access • Controlled access • Not sure

Q20 If data has controlled access: Who do you think should give permission for data to be shared and used again?	• The participants who took part should decide • The researcher(s) who collected it • The organisation where the original researcher(s) work • An independent committee • Other • Not sure	For each statement of who should give permission: Yes OR No • No Not sure
Q21 Who do you think should 'own' the data collected during a study?	• Me/the participants who took part • The researcher(s) who collected it • The organisation where the original researcher(s) work • Anyone who uses it • Whoever stores it • No one • Other • Not sure	For each statement of who owns it: Yes OR No • No • Not sure
Q22 Do you think it is important that researchers using shared data give feedback telling participants how their study data was used?	Yes No Not sure	Yes • Yes No • No • Not sure

Appendix J. Significant independent variable cross tabulations

Significant results highlighted in yellow.

	Q24 Age	Q23 Gender	Q25 Ethnicity	Q26 Education	Q27 Health rating	Deprivation Quintile	Source (Study)	Q3a Experience of taking part
Q24 Age		Pr = <0.001	Pr = <0.001	Pr = <0.001	Pr = <0.001	Pr = 0.007	Pr =< 0.001	Pr = <0.001
Q23 Gender	Pr = <0.001		Pr = <0.001	Pr = 0.055	Pr = <0.001	Pr = 0.470	Pr = <0.001	Pr = 0.008
Q25 Ethnicity	Pr = <0.001	Pr = <0.001		Pr = 0.029	Pr = <0.001	Pr = 0.189	Pr = <0.001	Pr = 1.000
Q26 Education	Pr = <0.001	Pr = 0.055	Pr = 0.029		Pr = <0.001	Pr = 0.548	Pr = <0.001	Pr = 0.920
Q27 Health rating	Pr = <0.001	Pr = <0.001	Pr = <0.001	Pr = <0.001		Pr = 0.342	Pr = <0.001	Pr = <0.001
Deprivation Quintile	Pr = 0.007	0.470	Pr = 0.189	Pr = 0.548	Pr = 0.342		Pr = <0.001	Pr = 0.168
Source (Study)	Pr = <0.001	Pr = <0.001	Pr = <0.001	Pr = <0.001	Pr = <0.001	Pr = <0.001		Pr = <0.001
Q3a Experience of taking part	Pr = <0.001	Pr = 0.008	Pr = 1.000	Pr = 0.920	Pr =< 0.001	Pr = 0.168	Pr = <0.001	

Appendix K. Significant dependent variable cross tabulations- add from separate file

	Q5	Q6_1	Q6_2	Q6_3	Q6_4	Q6_5	Q6_6	Q6_7	Q6_8
Q24 Age	Pearson chi2(3) = 33.8 Pr = <0.001								
Q23 Gender			Pearson chi2(2) = 22.1 Pr = <0.001						
Q25 Ethnicity		Pearson chi2(5) = 14.3 Pr = 0.014							
Q26 Education			Pearson chi2(6) = 29.2 Pr = <0.001			Pearson chi2(6) = 12.8 Pr = 0.047			Pearson chi2(6) = 22.4 Pr = 0.001
Q27 Health rating	Pearson chi2(5) = 14.2 Pr = 0.015		Pearson chi2(5) = 12.3 Pr = 0.031	Pearson chi2(5) = 29.3 Pr = <0.001	Pearson chi2(5) = 21.8 Pr = 0.001	Pearson chi2(5) = 16.9 Pr = 0.005	Pearson chi2(5) = 26.3 Pr = <0.001	Pearson chi2(5) = 17.4 Pr = 0.004	Pearson chi2(5) = 13.1 Pr = 0.022
Deprivation Quintile			Pearson chi2(4) = 13.2 Pr = 0.010						
Source (Study)	Pearson chi2(2) = 28.3 Pr = <0.001	Pearson chi2(2) = 13.9 Pr = 0.001	Pearson chi2(2) = 11.3 Pr = 0.004	Pearson chi2(2) = 12.2 Pr = 0.002	Pearson chi2(2) = 12.2 Pr = 0.002	Pearson chi2(2) = 7.2 Pr = 0.027	Pearson chi2(2) = 10.6 Pr = 0.005		
Q3a Experience of taking part			Pearson chi2(6) = 21.3 Pr = 0.002	Pearson chi2(6) = 17.3 Pr = 0.008		Pearson chi2(6) = 34.3 Pr = <0.001	Pearson chi2(6) = 32.1 Pr = <0.001	Pearson chi2(6) = 20.3 Pr = 0.002	Pearson chi2(6) = 16.6 Pr = 0.011

	Q7_1	Q7_2	Q7_3	Q7_4	Q7_5	Q7_6	Q7_7	Q7_8	Q7a
Q24 Age		Pearson chi2(3) = 12.6 Pr = 0.006	Pearson chi2(3) = 9.5 Pr = 0.023	Pearson chi2(3) = 27.6 Pr = <0.001	Pearson chi2(3) = 19.4 Pr = <0.001	Pearson chi2(3) = 60.9 Pr = <0.001	Pearson chi2(3) = 20.9 Pr = <0.001	Pearson chi2(3) = 49.6 Pr = <0.001	Pearson chi2(21) = 98.6 Pr = <0.001
Q23 Gender		Pearson chi2(2) = 13.8 Pr = 0.001		Pearson chi2(2) = 14.1 Pr = 0.001	Pearson chi2(2) = 11.6 Pr = 0.003		Pearson chi2(2) = 23.9 Pr = <0.001		Pearson chi2(14) = 49.1 Pr = <0.001
Q25 Ethnicity									
Q26 Education					Pearson chi2(6) = 14.2 Pr = 0.028	Pearson chi2(6) = 18.7 Pr = 0.005		Pearson chi2(6) = 12.9 Pr = 0.045	
Q27 Health rating						Pearson chi2(5) = 11.7 Pr = 0.039		Pearson chi2(5) = 11.6 Pr = 0.041	
Deprivation Quintile					Pearson chi2(4) = 11.4 Pr = 0.023				
Source (Study)		Pearson chi2(2) = 11.8 Pr = 0.003	Pearson chi2(2) = 7.8 Pr = 0.020	Pearson chi2(2) = 20.5 Pr = <0.001	Pearson chi2(2) = 20.7 Pr = <0.001	Pearson chi2(2) = 67.3 Pr = <0.001	Pearson chi2(2) = 23.1 Pr = <0.001	Pearson chi2(2) = 58.4 Pr = <0.001	Pearson chi2(14) = 72.6 Pr = <0.001
Q3a Experience of taking part			Pearson chi2(6) = 13.7 Pr = 0.033				Pearson chi2(6) = 24.0 Pr = 0.001		Pearson chi2(42) = 68.2 Pr = 0.006

	Q8_1	Q8_2	Q8_3	Q8_4	Q8_5	Q8_6
Q24 Age						Pearson chi2(3) = 9.6 Pr = 0.023
Q23 Gender			Pearson chi2(2) = 11.5 Pr = 0.003			
Q25 Ethnicity	Pearson chi2(5) = 18.1 Pr = 0.003					
Q26 Education	Pearson chi2(6) = 24.7 Pr = <0.001	Pearson chi2(6) = 13.2 Pr = 0.040			Pearson chi2(6) = 21.7 Pr = 0.001	
Q27 Health rating	Pearson chi2(5) = 18.3 Pr = 0.003	Pearson chi2(5) = 11.8 Pr = 0.037		Pearson chi2(5) = 19.7 Pr = 0.001	Pearson chi2(5) = 12.2 Pr = 0.033	
Deprivation Quintile			Pearson chi2(4) = 11.3 Pr = 0.023			
Source (Study)						Pearson chi2(2) = 8.9 Pr = 0.011
Q3a Experience of taking part	Pearson chi2(6) = 32.3 Pr = <0.001	Pearson chi2(6) = 17.9 Pr = 0.006	Pearson chi2(6) = 20.9 Pr = 0.002	Pearson chi2(6) = 29.8 Pr = <0.001	Pearson chi2(6) = 30.8 Pr = <0.001	Pearson chi2(6) = 28.8 Pr = <0.001

	Q9_1	Q9_2	Q9_3	Q9_4	Q9_5
Q24 Age	Pearson chi2(3) = 36.4 Pr = <0.001	Pearson chi2(3) = 120.0 Pr = <0.001	Pearson chi2(3) = 53.5 Pr = <0.001	Pearson chi2(3) = 24.5 Pr = <0.001	Pearson chi2(3) = 40.7 Pr = <0.001
Q23 Gender		Pearson chi2(2) = 7.6 Pr = 0.023			
Q25 Ethnicity					
Q26 Education	Pearson chi2(6) = 16.5 Pr = 0.011			Pearson chi2(6) = 14.7 Pr = 0.022	
Q27 Health rating	Pearson chi2(5) = 15.3 Pr = 0.009	Pearson chi2(5) = 16.2 Pr = 0.006	Pearson chi2(5) = 12.9 Pr = 0.024		
Deprivation Quintile		Pearson chi2(4) = 15.1 Pr = 0.005	Pearson chi2(4) = 11.7 Pr = 0.019	Pearson chi2(4) = 11.1 Pr = 0.026	
Source (Study)	Pearson chi2(2) = 73.2 Pr = <0.001	Pearson chi2(2) = 196.6 Pr = <0.001	Pearson chi2(2) = 91.7 Pr = <0.001	Pearson chi2(2) = 48.6 Pr = <0.001	Pearson chi2(2) = 75.8 Pr = <0.001
Q3a Experience of taking part	Pearson chi2(6) = 17.8 Pr = 0.007	Pearson chi2(6) = 13.1 Pr = 0.042	Pearson chi2(6) = 15.3 Pr = 0.018	Pearson chi2(6) = 30.7 Pr = <0.001	Pearson chi2(6) = 19.6 Pr = 0.003

	Q10_1	Q10_2	Q10_3	Q10_4	Q10_5
Q24 Age				Pearson chi2(6) = 167.9 Pr = <0.001	Pearson chi2(3) = 8.4 Pr = 0.038
Q23 Gender		Pearson chi2(2) = 10.5 Pr = 0.005	Pearson chi2(2) = 12.8 Pr = 0.002	Pearson chi2(4) = 23.1 Pr = <0.001	Pearson chi2(2) = 9.8 Pr = 0.008
Q25 Ethnicity					Pearson chi2(5) = 11.4 Pr = 0.043
Q26 Education	Pearson chi2(6) = 31.2 Pr = <0.001			Pearson chi2(12) = 31.4 Pr = 0.002	
Q27 Health rating				Pearson chi2(10) = 36.4 Pr = <0.001	
Deprivation Quintile					
Source (Study)				Pearson chi2(4) = 174.7 Pr = <0.001	
Q3a Experience of taking part				Pearson chi2(12) = 21.6 Pr = 0.043	Pearson chi2(6) = 27.3 Pr = <0.001

	Q11_1	Q11_2	Q11_3	Q11_4	Q11_5	Q11_6	Q11_7
Q24 Age			Pearson chi2(3) = 8.7 Pr = 0.034	Pearson chi2(3) = 37.9 Pr = <0.001		Pearson chi2(3) = 19.9 Pr = <0.001	Pearson chi2(3) = 13.2 Pr = 0.004
Q23 Gender		Pearson chi2(2) = 16.6 Pr = <0.001			Pearson chi2(2) = 10.0 Pr = 0.007		
Q25 Ethnicity							
Q26 Education							
Q27 Health rating					Pearson chi2(5) = 25.3 Pr = <0.001	Pearson chi2(5) = 11.9 Pr = 0.036	
Deprivation Quintile							
Source (Study)			Pearson chi2(2) = 6.2 Pr = 0.044	Pearson chi2(2) = 35.4 Pr = <0.001	Pearson chi2(2) = 11.8 Pr = 0.003	Pearson chi2(2) = 12.8 Pr = 0.002	
Q3a Experience of taking part			Pearson chi2(6) = 18.4 Pr = 0.005	Pearson chi2(6) = 14.9 Pr = 0.021	Pearson chi2(6) = 13.2 Pr = 0.040	Pearson chi2(6) = 17.7 Pr = 0.007	

	Q11_9	Q11_10	Q11_11	Q11_12	Q11_13	Q11_14	Q11_15
Q24 Age		Pearson chi2(3) = 15.5 Pr = 0.001		Pearson chi2(3) = 17.9 Pr = <0.001	Pearson chi2(3) = 10.6 Pr = 0.014		Pearson chi2(3) = 8.8 Pr = 0.032
Q23 Gender						Pearson chi2(2) = 6.7 Pr = 0.036	
Q25 Ethnicity							
Q26 Education							
Q27 Health rating							
Deprivation Quintile							
Source (Study)		Pearson chi2(2) = 15.4 Pr = <0.001		Pearson chi2(2) = 12.0 Pr = 0.002			Pearson chi2(2) = 7.4 Pr = 0.025
Q3a Experience of taking part	Pearson chi2(6) = 28.2 Pr = <0.001	Pearson chi2(6) = 21.6 Pr = 0.001	Pearson chi2(6) = 20.8 Pr = 0.002	Pearson chi2(6) = 25.6 Pr = <0.001			

	Q12_1	Q12_2	Q12_3	Q12_4	Q12_5
Q24 Age		Pearson chi2(3) = 13.9 Pr = 0.003		Pearson chi2(3) = 15.4 Pr = 0.002	Pearson chi2(3) = 88.8 Pr = <0.001
Q23 Gender		Pearson chi2(2) = 9.6 Pr = 0.008	Pearson chi2(2) = 6.1 Pr = 0.047	Pearson chi2(2) = 9.1 Pr = 0.010	
Q25 Ethnicity					
Q26 Education	Pearson chi2(6) = 16.5 Pr = 0.011				Pearson chi2(6) = 25.1 Pr = <0.001
Q27 Health rating	Pearson chi2(5) = 15.1 Pr = 0.010			Pearson chi2(5) = 21.9 Pr = 0.001	
Deprivation Quintile					
Source (Study)	Pearson chi2(2) = 6.5 Pr = 0.040	Pearson chi2(2) = 12.5 Pr = 0.002		Pearson chi2(2) = 16.6 Pr = <0.001	Pearson chi2(2) = 80.1 Pr = <0.001
Q3a Experience of taking part	Pearson chi2(6) = 15.3 Pr = 0.018	Pearson chi2(6) = 17.8 Pr = 0.007			Pearson chi2(6) = 93.8 Pr = <0.001

	Q13_1	Q13_2	Q13_3	Q13_4	Q13_5	Q13_6	Q13a
Q24 Age		Pearson chi2(3) = 14.8 Pr = 0.002				Pearson chi2(3) = 14.1 Pr = 0.003	
Q23 Gender		Pearson chi2(2) = 11.5 Pr = 0.003	Pearson chi2(2) = 11.5 Pr = 0.003	Pearson chi2(2) = 6.5 Pr = 0.038		Pearson chi2(2) = 22.1 Pr = <0.001	
Q25 Ethnicity							
Q26 Education	Pearson chi2(6) = 26.3 Pr = <0.001						Pearson chi2(6) = 20.8 Pr = 0.002
Q27 Health rating						Pearson chi2(5) = 11.6 Pr = 0.041	
Deprivation Quintile				Pearson chi2(4) = 12.0 Pr = 0.017		Pearson chi2(4) = 14.3 Pr = 0.006	
Source (Study)		Pearson chi2(2) = 15.9 Pr = <0.001				Pearson chi2(2) = 12.7 Pr = 0.002	
Q3a Experience of taking part	Pearson chi2(6) = 21.6 Pr = 0.001	Pearson chi2(6) = 20.8 Pr = 0.002				Pearson chi2(6) = 21.9 Pr = 0.001	

	Q14	Q15	Q16	Q17	Q18	Q19
Q24 Age		Pearson chi2(3) = 11.9 Pr = 0.007				
Q23 Gender	Pearson chi2(2) = 16.7 Pr = <0.001					
Q25 Ethnicity						
Q26 Education	Pearson chi2(6) = 30.3 Pr = <0.001					
Q27 Health rating		Pearson chi2(5) = 16.8 Pr = 0.005		Pearson chi2(5) = 13.8 Pr = 0.017	Pearson chi2(5) = 19.5 Pr = 0.002	
Deprivation Quintile					Pearson chi2(4) = 10.7 Pr = 0.030	
Source (Study)		Pearson chi2(2) = 7.8 Pr = 0.020				
Q3a Experience of taking part	Pearson chi2(6) = 17.9 Pr = 0.007	Pearson chi2(6) = 46.9 Pr = <0.001	Pearson chi2(6) = 14.2 Pr = 0.027	Pearson chi2(6) = 13.5 Pr = 0.036	Pearson chi2(6) = 38.5 Pr = <0.001	

	Q20_1	Q20_2	Q20_3	Q20_4	Q20_5
Q24 Age			Pearson chi2(3) = 20.9 Pr = <0.001		
Q23 Gender		Pearson chi2(2) = 9.1 Pr = 0.011	Pearson chi2(2) = 14.7 Pr = 0.001	Pearson chi2(2) = 9.6 Pr = 0.008	
Q25 Ethnicity					
Q26 Education				Pearson chi2(6) = 18.4 Pr = 0.005	
Q27 Health rating					
Deprivation Quintile					
Source (Study)			Pearson chi2(2) = 6.6 Pr = 0.036		Pearson chi2(2) = 6.9 Pr = 0.032
Q3a Experience of taking part					

	Q21_1	Q21_2	Q21_3	Q21_5	Q21_6	Q21_7	Q21_8	Q22
Q24 Age	Pearson chi2(3) = 17.1 Pr = 0.001	Pearson chi2(3) = 72.9 Pr = <0.001	Pearson chi2(3) = 10.1 Pr = 0.018					Pearson chi2(3) = 12.3 Pr = 0.006
Q23 Gender		Pearson chi2(2) = 14.5 Pr = 0.001		Pearson chi2(2) = 8.2 Pr = 0.016	Pearson chi2(2) = 29.3 Pr = <0.001			
Q25 Ethnicity						Pearson chi2(5) = 48.1 Pr = <0.001		
Q26 Education	Pearson chi2(6) = 14.4 Pr = 0.025	Pearson chi2(6) = 13.4 Pr = 0.037			Pearson chi2(6) = 18.2 Pr = 0.006			
Q27 Health rating								
Deprivation Quintile	Pearson chi2(4) = 13.6 Pr = 0.009	Pearson chi2(4) = 15.8 Pr = 0.003						
Source (Study)	Pearson chi2(2) = 38.9 Pr = <0.001	Pearson chi2(2) = 75.7 Pr = <0.001	Pearson chi2(2) = 37.5 Pr = <0.001				Pearson chi2(2) = 8.7 Pr = 0.013	Pearson chi2(2) = 14.4 Pr = 0.001
Q3a Experience of taking part		Pearson chi2(6) = 14.2 Pr = 0.027	Pearson chi2(6) = 13.1 Pr = 0.041					

Appendix L. Significant results after Bonferroni correction

Independent variable	Number of contrasts (i.e. pairwise comparisons)	Corrected p-value
Age	6	0.0083
Gender	3	0.0166
Experience of taking part	21	0.0023
Health rating	15	0.0033
Source study	4	0.0125

Question	Sub question	Independent Variable	Categories	% Split	Corrected P-value
Question 5 How concerned would you be if you knew data from the study that you are involved in was being shared?		Age	25-44 vs 45-64	60.4 vs 42.2	0.007
			25-44 vs 65-74	60.4 vs 44.1	<0.001
		Source study	ACONF vs ALSPAC	45.5 vs 60.5	<0.001
Question 6 How concerned would you be if you knew data was being shared with:	6b	Experience of taking part	Very positive vs not sure	19.8 vs 66.7	0.018
			Positive vs not sure	22.7 vs 66.7	0.036
	6c	Health rating	Excellent vs not sure	9.3 vs 66.7	<0.001
			Good vs not sure	8.0 vs 66.7	<0.001
			Average vs not sure	12.2 vs 66.7	<0.001
			Poor vs not sure	15.5 vs 66.7	0.001
			Very poor vs not sure	15.4 vs 66.7	0.006
		Experience of taking part	Very positive vs neither	7.1 vs 14.4	0.044
	6d	Health rating	Excellent vs not sure	5.6 vs 50.0	0.001
			Good vs not sure	6.9 vs 50.0	0.001
			Average vs not sure	9.5 vs 50.0	0.002
			Poor vs not sure	10.3 vs 50.0	0.006
	6e	Health rating	Excellent vs very poor	39.6 vs 84.6	0.017
			Average vs very poor	43.3 vs 84.6	0.046

Question	Sub question	Independent Variable	Categories	% Split	Corrected P-value
		Experience of taking part	Very positive vs positive	33.8 vs 43.1	0.023
			Very positive vs neither	33.8 vs 54.7	<0.001
	6f	Health rating	Excellent vs average	26.8 vs 38.5	0.024
			Excellent vs very poor	26.8 vs 69.2	0.017
			Excellent vs not sure	26.8 vs 83.3	0.044
			Good vs very poor	30.4 vs 69.2	0.040
		Experience of taking part	Very positive vs neither	25.3 vs 42.3	<0.001
			Very positive vs not sure	25.3 vs 77.8	0.014
	6g	Health rating	Excellent vs average	44.5 vs 58.0	0.010
		Experience of taking part	Very positive vs positive	46.1 vs 56.4	0.008
			Very positive vs neither	46.1 vs 60.2	0.010
	6h	Health rating	Excellent vs not sure	36.7 vs 100	0.026
Question 7 If data from the study in which you were involved was being shared, how concerned would you be about the following?	7b	Gender	Male vs female	75.1 vs 82.4	0.002
		Age	25-44 vs 65-74	77.6 vs 85.4	0.023
	7c	Age	25-44 vs 65-74	90.8 vs 95.7	0.047
	7d	Gender	Male vs female	81.7 vs 88.6	0.001
		Age	25-44 vs 65-74	88.6 vs 80.4	0.002
			25-44 vs 75-84	88.6 vs 50.0	0.002
			65-74 vs 75-84	80.4 vs 50.0	0.035
	7e	Gender	Male vs female	78.6 vs 84.4	0.016
		Age	25-44 vs 45- 64	79.7 vs 91.7	0.033
			25-44 vs 65-74	79.7 vs 89.3	0.001
	7f	Age	25-44 vs 45- 64	75.8 vs 94.1	<0.001
			25-44 vs 65-74	75.8 vs 94.6	<0.001
	7g	Gender	Male vs female	72.8 vs 83.5	<0.001
		Age	25-44 vs 65-74	76.9 vs 87.5	0.001
	7h	Age	25-44 vs 45- 64	70.5 vs 90.5	<0.001
			25-44 vs 65-74	70.5 vs 87.5	<0.001
Question 8 How likely would you be to give permission for your data to be shared for the following reasons?	8a	Health rating	Good vs average	94.5 vs 88.3	0.015
		Experience of taking part	Very positive vs neither	95.9 vs 85.4	<0.001
			Positive vs neither	92.7 vs 85.4	0.010
	8c	Experience of taking part	Very positive vs positive	81.4 vs 74.3	0.027
	8d	Health rating	Excellent vs average	80.2 vs 67.6	0.004

Question	Sub question	Independent Variable	Categories	% Split	Corrected P-value
		Experience of taking part	Very positive vs positive	81.2 vs 73.3	0.031
			Very positive vs neither	81.2 vs 65.8	<0.001
	8e	Health rating	Excellent vs average	88.0 vs 79.3	0.044
		Experience of taking part	Very positive vs positive	89.1 vs 82.5	0.036
			Very positive vs neither	89.1 vs 73.2	<0.001
			Positive vs neither	82.5 vs 73.2	0.044
	8f	Experience of taking part	Very positive vs positive	75.5 vs 65.7	0.006
			Very positive vs neither	75.5 vs 57.6	<0.001
Question 12 How and when would you like to be asked to share your data?	12b	Gender	Male vs female	36.5 vs 44.5	0.008
		Age	25-44 vs 65-74	43.8 vs 33.6	0.011
		Experience of taking part	Very positive vs neither	34.7 vs 48.4	0.008
	12d	Gender	Male vs female	7.8 vs 4.2	0.011
		Age	25-44 vs 65-74	4.0 vs 9.2	0.003
	12e	Age	25-44 vs 45-64	0.0 vs 4.8	0.009
			25-44 vs 65-74	0.0 vs 8.1	<0.001
		Experience of taking part	Very positive vs negative	0.9 vs 50.0	<0.001
			Very positive vs not applicable	0.9 vs 16.7	0.006
			Positive vs negative	0.8 vs 50.0	<0.001
			Positive vs not applicable	0.8 vs 16.7	0.006
			Neither vs negative	1.9 vs 50.0	<0.001
			Neither vs not applicable	1.9 vs 16.7	0.016
			Negative vs very negative	50.0 vs 0.0	0.001
			Negative vs not applicable	50.0 vs 16.7	<0.001
			Negative vs not sure	50.0 vs 0.0	<0.001
Question 13 What information would you like to see on the consent form before you agree to share your data?	13a	Experience of taking part	Neither vs negative	99.0 vs 75.0	0.041
			Neither vs not applicable	99.0 vs 75.0	0.041
	13b	Gender	Male vs female	91.3 vs 95.5	0.003
	13b	Experience of taking part	Very positive vs negative	94.6 vs 50.0	0.001
			Positive vs negative	94.8 vs 50.0	0.001
			Neither vs negative	96.4 vs 50.0	0.001
	13c	Gender	Male vs female	87.6 vs 92.3	0.009
	13d	Gender	Male vs female	71.9 vs 77.9	0.032
	13f	Gender	Male vs female	90.7 vs 96.5	<0.001

Question	Sub question	Independent Variable	Categories	% Split	Corrected P-value
		Experience of taking part	Very positive vs negative	94.3 vs 50.0	0.001
			Positive vs negative	96.2 vs 50.0	<0.001
			Neither vs negative	96.4 vs 50.0	<0.001
			Negative vs not applicable	50.0 vs 100	0.019
			Negative vs not sure	50.0 vs 88.9	0.049
Question 15 Does knowing about data sharing affect the likelihood of respondents taking part in research?		Age	25-44 vs 65-74	50.5 vs 60.4	0.017
		Health rating	Excellent vs average	58.4 vs 46.7	0.047
		Experience of taking part	Very positive vs positive	62.4 vs 46.2	<0.001
			Very positive vs neither	62.4 vs 39.4	<0.001
Question 20 If data has controlled access: Who do you think should give permission for data to be shared and used again?	20b	Gender	Male vs female	14.9 vs 21.3	0.008
	20c	Gender	Male vs female	28.9 vs 20.2	<0.001
	20d	Gender	Male vs other	10.6 vs 30.8	0.039
			Female vs other	8.2 vs 30.8	0.016
	20e	Source study	ACONF vs ALSPAC	5.3 vs 9.5	0.026
Question 21 Who do you think should 'own' the data collected during a study?	21a	Age	25-44 vs 75-84	55.0 vs 10.0	0.026
			45-64 vs 65-74	41.7 vs 60.1	0.017
			65-74 vs 75-84	60.1 vs 10.0	0.010
		Source study	ACONF vs participant groups	52.7 vs 11.1	<0.001
			ALSPAC vs participant groups	50.2 vs 11.1	<0.001
	21b	Age	25-44 vs 45-64	55.9 vs 28.6	<0.001
			25-44 vs 65-74	55.9 vs 33.2	<0.001
			25-44 vs 75-84	55.9 vs 0.0	0.002
		Source study	ACONF vs participant groups	33.7 vs 6.4	<0.001
			ALSPAC vs participant groups	50.9 vs 6.4	<0.001
			ACONF vs ALSPAC	33.7 vs 50.9	<0.001
	21c	Age	25-44 vs 75-84	48.7 vs 0.0	0.013
			45-64 vs 75-84	46.4 vs 0.0	0.033
			65-74 vs 75-84	45.6 vs 0.0	0.027
		Source study	ACONF vs participant groups	49.1 vs 7.9	<0.001
			ALSPAC vs participant groups	44.3 vs 7.9	<0.001
Question 22 Do you think it is important that researchers using shared data give feedback telling participants how their study data was used?		Age	25-44 vs 45-64	83.8 vs 71.4	0.027
		Source study	ACONF vs ALSPAC	74.9 vs 83.7	0.001

Blank page