The evaluation and harmonisation of disparate information metamodels in support of epidemiological and public health research

Christiana McMahon

Farr Institute of Health Informatics Research
Institute of Health Informatics
UCL

A thesis presented for the award of Doctor of Philosophy

Declaration

I, Christiana McMahon confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Background: Descriptions of data, metadata, provide researchers with the contextual information they need to achieve research goals. Metadata enable data discovery, sharing and reuse, and are fundamental to managing data across the research data lifecycle. However, challenges associated with data discoverability negatively impact on the extent to which these data are known by the wider research community. This, when combined with a lack of quality assessment frameworks and limited awareness of the implications associated with poor quality metadata, are hampering the way in which epidemiological and public health research data are documented and repurposed. Furthermore, the absence of enduring metadata management models to capture consent for record linkage metadata in longitudinal studies can hinder researchers from establishing standardised descriptions of consent.

Aim: To examine how metadata management models can be applied to ameliorate the use of research data within the context of epidemiological and public health research.

Methods: A combination of systematic literature reviews, online surveys and qualitative data analyses were used to investigate the current state of the art, identify current perceived challenges and inform creation and evaluation of the models.

Results: There are three components to this thesis: a) enhancing data discoverability; b) improving metadata quality assessment; and c) improving the capture of consent for record linkage metadata. First, three models were examined to enhance research data discoverability: data publications, linked data on the World Wide Web and development of an online public health portal. Second, a novel framework to assess epidemiological and public health metadata quality framework was created and evaluated. Third, a novel metadata management model to improve capture of consent for record linkage metadata was created and evaluated.

Conclusions: Findings from these studies have contributed to a set of recommendations for change in research data management policy and practice to enhance stakeholders' research environment.

Acknowledgements

Over the past four years, I have had the privilege of working alongside a group of incredible people without whom completing this thesis would not have been possible.

- *To Dr Spiros Denaxas,*

  Your endless support and wisdom guided me through my project; you taught me what it is to be an academic,

- *To Dr Henry Potts and Dr Bernard de Bono,*

  For your invaluable time and advice,

- *To Dr Tito Castillo and Professor Dennis Kehoe,*

  For your help and supervision,

- *To Professor Carol Dezateux,*

  For your irreplaceable guidance during the MPhil stage of my project,

- *To the participants of the data discoverability and metadata quality studies, and the longitudinal studies,*

  For your kindness; your collective sense of selflessness enabled me to undertake and complete my studies,

- *To the Data Lab team, Clinical Epidemiology Group, and friends across the Farr Institute of Health Informatics Research,*

  For your key insight and boundless enthusiasm,

- *To the epiLab team and colleagues at the UCL Institute of Child Health,*

  For your vital assistance,

- *To all my family,*

  For being my support network and encouraging me every step of the way,

and to so many others,
thank you.

This thesis was funded by the Medical Research Council as a 4year CASE award with AIMES Grid Services CIC. The data discoverability study was commissioned by the Wellcome Trust on behalf of the Public Health Research Data Forum. I was also in receipt of funding from the UCL Institute of Health Informatics.

Dedication

I would like to dedicate this thesis to my parents,
Androulla & Ian McMahon.

Table of contents

List of tables

List of figures

List of supplementary tables

List of supplementary figures

List of supplementary codes

Publications and conference presentations

Publications

McMahon, C., S. Denaxas. (2017). “A novel metadata management model to capture consent for record linkage in longitudinal studies”. Informatics for Health and Social Care. *in press*.

McMahon C, Denaxas S. “A novel framework for assessing metadata quality in epidemiological and public health research settings.” *AMIA Summits on Translational Science Proceedings*. 2016;2016:199-208.

McMahon, C., T. Castillo, et al. (2015). "Improving metadata quality assessment in public health and epidemiology." Stud Health Technol Inform **210**: 939.

Castillo, T., A. Gregory, S. Moore, B. Hole, **C. McMahon**, S. Denaxas, V. Van den Eynden, H. L’Hours, L. Bell, J. Kneeshaw, M. Woollard, C. Kanjala, G. Knight, B. Zaba. (2014). “Enhancing Discoverability of Public Health and Epidemiology Research Data”. Wellcome Trust, United Kingdom.

Conferences: Oral presentations

McMahon C., and S Denaxas. A novel framework for assessing metadata quality in epidemiological and public health research settings. Presentation at: AMIA Summits on Translational Science; 2016 March 21-23; San Francisco, USA

McMahon C., T Castillo, et al. Improving the capture of consent for record linkage metadata in UK longitudinal studies. Presentation at: The Farr Institute PhD Symposium; 2015 June 9-10; Manchester, United Kingdom

McMahon, C., V. Van den Eynden and B Hole. Enhancing the discoverability of public health and epidemiology research data. Presentation at: Public Health Research Data Forum Meeting. 2014 January 20-21; Wellcome Trust, London, United Kingdom

Conferences: Poster presentations

McMahon, C., S. Denaxas, et. al. (2017). "Methods for enhancing biomedical research data discoverability". Poster presented at: Informatics for Health; 2017 April 24-26; Manchester, United Kingdom

McMahon, C., S. Denaxas, et al. Public health and epidemiology metadata: Informing development of a quality assessment framework. Poster presented at: The Farr Institute International Conference 2015 Data Intensive Health Research and Care; 2015 August 26-28; St Andrews, Scotland, United Kingdom

McMahon, C., T. Castillo T, et al. Improving metadata quality assessment in public health and epidemiology. Poster presented at: Medical Informatics Europe Digital healthcare empowering Europeans. 2015 May 27-29; Madrid, Spain

McMahon, C., C. Dezateux, et al. Development of a novel metadata format to record longitudinal study consent models for record linkage. Poster presented at: Exploring Existing Data for Health Research. 4th SHIP Conference; 2013a August 28-30; St Andrews, Scotland, United Kingdom

McMahon, C., C. Dezateux, et al. Longitudinal studies and the research data lifecycle: Application of the Data Documentation Initiative. Poster presented at: Infrastructure, Intelligence, Innovation: driving the Data Science agenda. 8th International Digital Curation Conference; 2013b January 14-17; Amsterdam, The Netherlands http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster207.pdf

Achievements and public engagement

AMIA Summits on Translational Science: I was a finalist in the best student paper competition for my work on a novel framework to assess metadata quality in epidemiological and public health research settings. I presented my work at the conference in San Francisco, USA (2016) and utilised this opportunity to engage with other researchers and stakeholders involved in research focusing on metadata quality. I was responsible for all aspects of this work.

Biomedical Research Infrastructure Software Service Community Meet and Hack event: I was part of a team which was awarded 1st place for work looking at the use of i2b2 (ontology management software) and ICD codes to identify patients with a particular health condition from an exemplar Hospital Episode Statistics (HES) dataset. The work was completed and presented at the Biomedical Research Infrastructure Software Service (BRISSKit) Community Meet and Hack event hosted at the University of Leicester, 9th -11th October 2012. I was responsible for demonstrating the tool and helping to answer questions.

Engaging with the public and prospective students: I presented different aspects of my thesis at open events held at UCL Institute of Child Health (2012) and UCL Institute of Health Informatics (2015). These events were designed to engage with the public and in particular potential students. The posters helped to facilitate discussion around my Ph.D. project and the implications of my research within the context of public health and epidemiological research data management.

Chapter 1 Introduction

1.1 Epidemiological and public health research studies

Epidemiological and public health data hold considerable research potential. These data provide a richness and longevity that can be harnessed to investigate disease and inform health policy and practice. Bringing together these data to enable these applications require increased standardisation and improved documentation.

Electronic health records (EHRs) - longitudinal records of health and administration of care for an individual, potentially spanning several healthcare-related organisations (BS EN ISO 2012) - contain a set of rich data which can be utilised for secondary purposes such as research. By generating clinically phenotyped cohorts based on information sourced from EHRs, researchers are able to investigate disease across large populations at a relatively low cost (Rubbo, Fitzpatrick et al. 2015). Clinical information stored in EHRs has become a valuable resource and their strategic secondary use is helping to maximise their research benefits (Martin 2003; Coveney 2005; Chute, Ullman-Cullere et al. 2013). There are a variety of different data types, Table 1-1.

Table 1-1 Types of data

Type of data	Description
Clinical/General	These records contain general clinical information about the administration of healthcare. These records can be produced and maintained at GP surgeries as part of primary care or hospitals as part of the administration of secondary care.
Genetic	Following genetic screening and diagnostic tests in secondary care, genetic information is collated and stored in the clinical records of patients.
Imaging	Images produced from MRI, PET or ultrasound scans are often produced in secondary care settings such as hospitals.
Audiological	These records assist with the treatment and/or managing of audiological conditions including the enabling of certain physiological and psychological interventions.

National resources such as the UK Biobank have to date recruited 500,000 participants aged 40-69 from 2006 to 2010 (2016p). Consent was requested from the participants to gather anthropometric measurements and obtain biological samples. Resources such as these provide researchers with a comprehensive set of data which may be investigated; findings from analyses can help to inform recommendations for change in health policy and practice.

1.1.1 Record sharing and record linkage

In the UK and other countries such as the USA, Sweden, and Australia, it is possible to share and link records together. Record sharing involves giving a researcher or research group access to particular records. Record linkage involves combining records together from disparate sources to create enriched datasets. The research potential of these datasets can be harnessed as part of more complex investigations into the origins of disease. There are two types of record linkage: a) deterministic record linkage involves integrating records using unique common identifiers such as the NHS number (a unique series of ten digits from zero to nine assigned at birth or at first interaction with the healthcare system which uniquely identifies a patient) from multiple disparate health datasets; and b) probabilistic record linkage involving stakeholders calculating the probability that the records related to the same individual. It is also possible to link health records to other types of administrative records such as education.

To perform record linkage in the UK, consent from research participants can be requested or permission to suspend the need for consent may be obtained from the National Information Governance Board. (Knies, Burton et al. 2012) The consent form itself must be written using accessible terms and encourage conversations between the participant and researcher. (Gori, Greco et al. 2012) Furthermore, the participant must be given the opportunity to ask questions and gain further clarification before confirmation of understanding and informed consent can be obtained.

1.1.2 Reusing existing datasets for epidemiological and public health research purposes

Epidemiological and public health research studies draw on the wealth of clinical data available from sources such as the Clinical Practice Research Datalink (CPRD) (2016f). These disparate data sources are linked and their research opportunities harnessed by research platforms such as CALIBER (Denaxas, George et al. 2012). CALIBER combines linked electronic health records from various sources including CPRD.

Sharing epidemiological and public health data facilitates collaboration between multiple institutions helping to achieve common research goals, reduce redundant efforts, and promote transparency (Thiru, Hassey et al. 2003; Tenopir, Allard et al. 2011; Borgman 2012). For example, research into Autism spectrum disorder (ASD) demands large quantities of genotypic and phenotypic data; therefore, cross-organisational collaborations are vital if results are to be replicated and the causes of disease better understood (Johnson, Whitney et al. 2010). In the case of muscular dystrophy, families often altruistically share their clinical data with researchers, knowing that the findings of such studies may not necessarily be of direct benefit to themselves (Kush and Goldman 2014). Alzheimer's research has also benefited from data sharing on a global scale. The Alzheimer's Disease Neuroimaging Initiative (ADNI) (2016a) provides researchers with MRI and PET images along with genetic, cognitive test and biomarkers to enable research into this area of mental health disease. Another example of where the sharing of data has enabled epidemiological and public health research is the Emerging Risk Factors Collaboration (ERFC) (2016h). This is a central database which brings together over 125 studies with over 2 million participants to facilitate cardiovascular research.

Further, sharing research data can potentially lead to an increase in citation rates. According to a study by Piwowar, Day et al. (2007) looking at the association between cancer microarray data sharing practices in clinical trials and increased citation rate, the authors found researchers who shared their data were cited approximately 70% more than those who did not.

On 10 January 2011, a Joint Statement of Purpose was launched to address issues relating to the availability of public health and realise visions of increased data sharing and efficiency of use (Walport and Brest 2011). Pisani, Whitworth et al. (2009) suggested that barriers to data sharing can be categorised into the following: a) ethical, b) technical, and c) professional. The authors also suggested that improved data management is vital particularly in certain developing countries where it is virtually non-existent.

Furthermore, limited academic and career incentives can discourage researchers from publishing datasets. Consequently, the potential for data sharing and recognition to be awarded to those involved in research data management is reduced (Pisani, Whitworth et al. 2009). Other challenges associated with data sharing as identified through a systematic literature review include: financial costs, concerns over effective data governance and insufficient technology to meet the demands of data sharing whilst maintain patient privacy and confidentiality (Hopf, Bond et al. 2014).

According to a study by Tenopir, Allard et al. (2011), limited data documentation (metadata) can also negatively impact the extent to which research data are shared. An inability to fully understand the data can potentially reduce scope for data reuse. There are also issues relating to the long-term archiving of such data and the undervaluing of the effort taken by those responsible for preparing the dataset for submission to a repository (Pisani and AbouZahr 2010; Kolker and Stewart 2014).

Furthermore, inconsistencies between steps taken to manage and curate research data can potentially reduce the extent to which these data may be combined and used as part of more complex queries (Kush and Goldman 2014). This then gives rise to other potential problems such as reduced accuracy in meta-analyses as researchers are unable to analyse certain datasets for potential inclusion in their study (Ioannidis 2012).

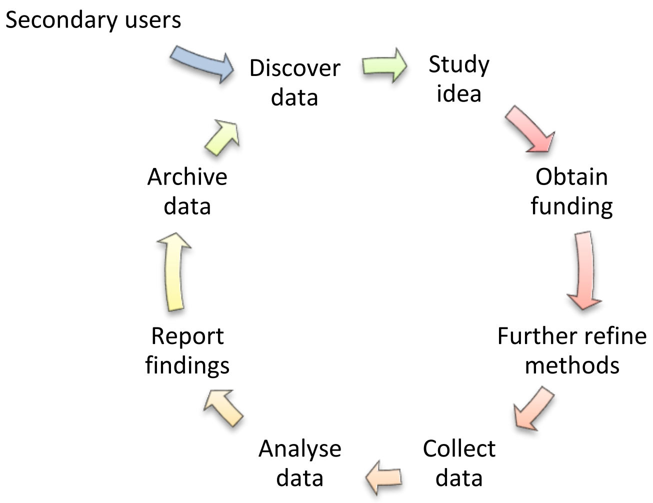
In epidemiological and public health research studies, the use of linked longitudinal data from multiple disparate clinical data sources must be matched with metadata management frameworks to facilitate meaningful use of these data (Mougin, Burgun et al. 2006; Safran, Bloomrosen et al. 2007). Such frameworks include those designed to support research data storage, interoperability and longer-term archiving of data and metadata (Meredith, Crouch et al. 2010).

1.1.3 A data lifecycle-based approach to epidemiological and public health research studies

The inherent complexity and heterogeneity of clinical data can lead to difficulties when trying to process these data for epidemiological and public health research. A motivation to promote a cyclical approach to research data management, involving increased data reuse and repurposing, has begun to catalyse a shift in research culture. Stakeholders are increasingly being encouraged to work even more closely to maximise the potential benefits associated with this cyclical approach to research data management.

The research data lifecycle (RDL) maps the different stages associated with a research study, Figure 1-1.

Figure 1-1 Stages of the research data lifecycle



Once a study has finished, findings have been published, and the data archived, there could be potential to repurpose the data and reuse it. Therefore, the life of that research data continues and a new cycle begins. The RDL successfully conveys the associated potential longevity of epidemiological and public health research data. It is through this process of data discovery, repurposing and reuse that the life of research data continues. Nevertheless, the extent of use is in accordance with consent given by the participants, stipulations from funding agencies, embargos, applicable law(s) and any other factors which could potentially affect use. It is also possible for data to be destroyed, known as data destruction, should this be necessary.

1.2 Drivers for change in epidemiological and public health research data management policy and practice

1.2.1 Changes in research culture

The agreeing of the FAIR principles for managing and stewarding scientific data signify a collective acknowledgement of the importance of enhanced research data management. The principles, Findability, Accessibility, Interoperability, and Reusability serve to guide researchers and other stakeholders on how to maximise potential research benefits associated with publications, tools and other such scholarly artefacts. Having FAIR research artefacts can help to address issues relating to the research data management such as, having sufficient metadata to describe a resource for discoverability and interoperability purposes (Wilkinson, Dumontier et al. 2016).

These changes in research culture also extend to the way in which scholarly publications are managed. As from April 2016, for a publication to be eligible for inclusion in the next Research Excellence Framework, (2016o) the author accepted manuscript must be made openly accessible within 90 days of acceptance. Also, the Policy on Open Access (RCUK 2013) stipulates that all publicly funded research papers must be openly accessible. These policies are also in keeping with recommendations made by the European Commission (2012) and World Health Organization (2013b) to enhance the research environment.

These changes in the way in which publications are managed encourages researchers to adapt the way in which they handle their research data in both the short and longer term. Often journals, such as PLOS (2016n), request the underlying research data be made accessible to the reviewer and sometimes request researchers deposit their data into repositories. Currently, there is a need for improved approaches to managing epidemiological and public health research data particularly given their increasing complexity and the potential for future re-use (Anderson, Lee et al. 2007).

1.2.2 Funder policies and increased investment

Funder policies have been a driving factor in the shift towards enhanced research data management practice; whilst safeguarding the privacy and confidentiality of research study participants. In 2015, the RCUK revised the Common Principles on Data Policy (RCUK 2015b) and provided guidance on best practice in the management of research data (RCUK 2015a). In 2016, the RCUK also published the Concordat on Open Research Data which sought to ensure research data are made openly available within the limits of any applicable legal, ethical and regulatory frameworks (RCUK 2016).

Certain funding agencies are requesting that researchers include data access statements in their journal article with funding councils such as the EPSRC stating that making data available solely upon request via email is insufficient. By researchers describing how the underlying research data can be accessed, the potential for data to be discovered, repurposed and reused potentially increases as does the scope for replicating and verifying research findings. There is also the potential to assign a globally unique, persistent identifier such as a DOI (digital object identifier) to the journal article and data and for these both to be citable.

Furthermore, in January 2013, it was concluded by fifty delegates from eight countries that to meet the requirements of the clinical, research and patient communities, a global alliance, and technology platforms with open standards are needed to support data sharing (2013a). Another such example is the agreement of the UK government to investigate secure methods of data sharing in support of research (Economic & Social Research Council 2013) based on the findings of a report published by the Administrative Data Taskforce (2012).

1.3 Metadata

1.3.1 Definition

A fundamental component to the cyclical use of research data is the provision of, and access to, data documentation – metadata.

Metadata are data about data and the process through which they were collected.

Research artefacts such as data dictionaries (an example of metadata), are indispensable to researchers and support secondary use of clinical data for research.

1.3.2 Stakeholders: creating and using metadata

Metadata can be created and used by stakeholders in epidemiological and public health research settings across the research data lifecycle. For example, researchers and other users of data can create metadata during the data collection phase of projects. Information about which data were collected, when and why, can form research artefacts such as data dictionaries which are an example of metadata. They can also create metadata when analysing these data in the form of notes contained within analysis plans. Researchers can also create metadata detailing how research data should be stored and accessed (e.g. access to data safe havens are needed to enact data access protocols) in the form of data management plans. Subsequently, data managers, librarians, archivists and other stakeholders involved in managing research data can utilise these metadata in the form of data management plans to store these data on both a short and longer term basis.

Other stakeholders in epidemiological and public health research, such as funders, can create and share metadata of the research studies to which they are affiliated. These metadata could be contained within online catalogues which can be utilised by other stakeholders; an example of this is the MRC Gateway. Collections of metadata records can help members of the research community to learn of existing research studies and which data were collected. Having access to this kind of information is particularly important when wanting to discover and reuse existing data to maximise their research potential.

1.4 Metadata and its importance to epidemiological and public health research

Epidemiological and public health data have many potential research benefits. For these benefits to be realised, and research opportunities maximised, the users of these data need access to good quality metadata. Good quality metadata are key to opening up epidemiological and public health data for potential repurposing and reuse. Metadata provide the much needed context to enable stakeholders to investigate these data further and determine where additional research benefits may be realised (Liolios, Schrimi et al. 2012).

1.4.1 The role of metadata in honouring consent and other ethical issues

Honouring the limits of consent in a research study is a key aspect of epidemiological and public health research. Metadata are important to helping to protect the rights of study participants and assisting stakeholders in observing the limits of consent and addressing other ethical issues. For example, having access to good quality metadata can help stakeholders to determine where the Data Protection Act 1998 applies; for example, the act would apply if the study involves living participants. Having access to this kind of metadata is important to helping stakeholders fulfil legal and ethical obligations when using these research data.

Access to good quality metadata can also help stakeholders to determine under which circumstances ethical approval was given to the primary researchers/data users. Having access to this kind of metadata is particularly important when wanting to share and reuse data ethically. In a study by de Vries, William et al. (2014), a concern shared by some of the interviewees around data sharing was that ethical approval was awarded to collect and use the data in a particular way. However, secondary users of the data could potentially use these data differently to the primary research team. Consequently, secondary users of the data must ensure the data are used in accordance with the participants' wishes and that legal and ethical obligations have been met. Hence, having access to good quality metadata detailing the scope of ethical approval is important to helping secondary users of data meet ethical obligations.

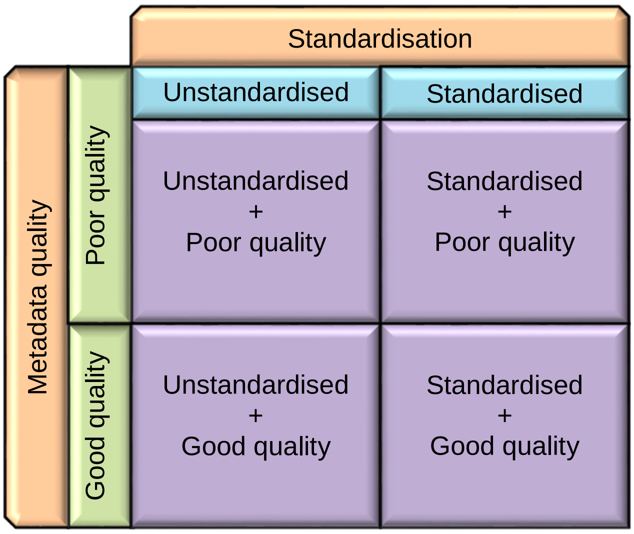
Furthermore, having access to good quality metadata can also enable stakeholders to begin charactering research data without having direct access. For example, having detailed metadata about variables could help data users to gain a better understanding of which data were collected and how they may be utilised. Withholding direct access to the research data until such time it can be made available, could help to protect the privacy and confidentiality of participants, particularly in studies where the data collected is of a highly sensitive, personally identifiable nature. It is in these situations where good quality metadata is pertinent to enabling the research process.

1.4.2 Challenges I: poor quality metadata

Harnessing the benefits associated with metadata is challenging given the current inconsistencies in the availability and quality of these metadata. For example, in genomics research, stakeholders wanting to contextualise genomic research data, are in some instances unable to do so due to a lack of available metadata (de Vries, Williams et al. 2014). There are also instances whereby metadata contain missing fields and inaccuracies (Panahiazar, Dumontier al. 2017; Dumontier, Gray et al. 2016; Marc, Beattie et al. 2016). Consequently, stakeholders wanting to discover and characterise data, with a view to repurposing and reusing them for research, are hindered from doing so. It is these challenges which give raise to the need for improved quality of metadata and the subsequent enhanced discoverability of research data.

Further, metadata are often not subject to the same level of the scrutiny as the research data to which they are associated. However, application of standards only affects the way in which the metadata elements are structured and does not necessarily influence the way in which the corresponding fields are completed by stakeholders. Therefore, the quality of metadata instances is variable; a potential outcome is standardised, poor quality metadata Figure 1-2.

Figure 1-2 High level metadata quality vs standardisation categories



		Standardisation	
		Unstandardised	Standardised
Metadata quality	Poor quality	Unstandardised + Poor quality	Standardised + Poor quality
	Good quality	Unstandardised + Good quality	Standardised + Good quality

Acknowledgement is given the varying degrees of compliance to standards and the quality of metadata and Figure 1-2 is designed to present a very high level view of how instances of metadata could potentially be categorised according to these two factors.

1.4.3 Challenges II: lifecycle-based metadata

According to a literature review by Ochoa and Duval (2009) two methods of metadata evaluation were identified: a) manual, which involves individually reviewing metadata and evaluating quality, and b) statistical - using computational techniques for defining metrics to evaluate metadata quality. Nonetheless, metadata are domain-specific and require fit-for-purpose metrics be created and evaluated to assess quality. Given that metadata are critical to the research process, much is needed in the way of developing and integrating metadata quality assessment into stakeholders' workflows in public health and epidemiological research and across the stages of the research data lifecycle.

At the data discovery and access stages, the provision of high quality metadata has the potential to help characterise the data and support researchers in deciding whether or not to request access (Taylor, Field et al. 2008; Pickett, Liu et al. 2013). According to a study by Rans, Day et al. (2013) evaluating current mechanisms for public health data citation of which metadata plays a key role, assigning unique and enduring identifiers to datasets, coupled with landing pages with high level metadata are needed to facilitate data discovery and access. The use of high quality metadata in these situations will help stakeholders to better cite datasets and indicate through which means they were accessed.

Furthermore, data sharing practices vary amongst researchers and access to good quality metadata inclusive of data models can help researchers to better understand the data (Li, Wen et al. 2012; Van den Eynden 2012). According to a study by Wang, Vergara-Niedermayr et al. (2014) looking at metadata-based management and sharing of distributed biomedical data, three mechanisms for sharing data exist: a) centralised location using a single schema; b) federated architecture; and c) distributed architecture. Metadata quality plays a fundamental role in enabling data sharing and reuse as they provide the contextual information needed to better researchers' understanding of how the data were collected, when, and how variables were managed. Another example of where the potential for data sharing and reuse can be hampered through the provision of poor quality metadata is that of managing repositories. According to a study by Neu, Crawford et al. (2012) looking at how heterogeneous neuroimaging metadata are managed by global repositories, the need to share usable metadata combined with the heterogeneity of research data are challenging aspects of research data management.

At the data integration and harmonisation stages of the RDL, harmonising phenotypic data can benefit from good quality, and where possible, standardised metadata. These processes are further supported through access to contextual information in the form of metadata. The metadata can signal the commonalities between disparate datasets and support the performance of meta-analyses (Vardaki, Papageorgiou et al. 2009; Davies, Gibbons et al. 2014). Nevertheless, the lack in uptake of standards coupled with inconsistent quality of metadata can render analyses of epidemiological and public health research data problematic (MRC 2014). Having a robust and standardised approach to metadata markup can help better support the documenting of research studies (Kolker, Ozdemir et al. 2014).

1.5 Research aim and objectives

1.5.1 Research aim

The overarching aim of my research was to create and evaluate a series of metadata management models to improve the reuse of epidemiological and public health data within research settings. The metadata management models support researchers and other stakeholders in harmonising data from disparate sources by providing standards-based mechanisms to help manage metadata across the research data lifecycle. These metadata management models build on existing standards and practices and consider how these can be reused and adapted to enhance the way in which researchers and other stakeholders manage their metadata.

1.5.2 Research question

How can information standards be applied at various points of the research data lifecycle to create information metamodels to facilitate the reuse and repurposing of research data within epidemiological and public health research?

1.5.3 PhD research objectives

I. To systematically review and evaluate methods for enhancing the discoverability of public health and epidemiological research data

II. To create and evaluate a novel quality assessment framework for epidemiological and public health metadata

III. To create and evaluate a novel metadata management model to support improved recording of consent for record linkage metadata in bespoke investigator-led cohort studies
To make a series of recommendations for improving the management of epidemiological and public health research data

The following case study illustrates the challenges this Ph.D. addresses and demonstrates the novelty of my research:

Case study: Millennium Cohort Study

The Millennium Cohort Study is a UK-based study which follows the lives of approximately 19000 children born between 2000 and 2001[1]. Birth cohort studies, in addition to other like studies, are an invaluable resource to researchers and other stakeholders wanting to investigate life course influences and origins of disease. The ability to reuse and repurpose research data from these studies is imperative to maximising resources and further investigating health conditions. Researchers are able to access the data from the Millennium Cohort Study through the UK Data Archive. Here, standardised metadata accompanies the data to facilitate its use.

Nevertheless, with heightened discoverability, and greater availability of standardised metadata, the potential for reuse could be further enhanced. The aim of this Ph.D. project will be to address these current challenges facing the research community by applying information standards to create information metamodels to facilitate the reuse and repurposing of research data within public health and epidemiological research settings.

The outcomes of the discoverability study (research case study I) will help to improve the current discoverability of this study and where associated research data may be found and potentially accessed. The outcomes of the metadata quality assessment study (research case study II) will help stakeholders to assess the quality of the metadata associated with this and other studies. The outcomes of the third study – recording consent to record linkage metadata (research case study III) will help to standardise and hence improve the way in which the associated consent forms are recorded.

1. http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=851&sitesectiontitle=Welcome+to+the+Millennium+Cohort+Study

1.5.4 Outline of research studies

The first component of this thesis examines how metadata-based mechanisms can enhance the discoverability of epidemiological and public health research data. I then build on this examination by investigating how metadata management models can improve the approach to assessing metadata quality within these research settings. The third component of this thesis explores how consent for record linkage metadata can be recorded using a series of novel metadata management models. The third component also demonstrates how the recording of such metadata can be standardised and critically appraises the current prevailing standard for application to epidemiological and public health research settings. The issue of discoverability was firstly addressed as awareness of existing research datasets is a key aspect to the reuse and repurposing of research data. Having established that an existing research dataset may lend itself to a future study, the next step is to characterise the data; this gave rise to the metadata quality study. Having good quality metadata is vitally important to gaining an understanding of the data and its potential research opportunities. Finally, when using the data from multiple research studies, it is important to understand which the limits of consent are. Having standardised metadata describing the consent process, is critical to helping to facilitate the reuse of research data whilst respecting the wishes and rights of the participants.

1.6 Researcher involvement

The following outlines my role and responsibilities for each research case study presented in this thesis:

1.6.1 Data discoverability study

I was responsible for:

- all aspects of the review, and designing and developing the survey,
- all quantitative and qualitative analyses performed and drawing of the graphs,

- identifying the need for a registration process for observational studies which could be enacted through a public health portal similar to that of ClinicalTrials.gov. It was through my systematic literature review I identified the potential to use data publications as another way of enhancing data discoverability. It was my research in this thesis, I identified the potential to use semantic web technologies as mechanisms to further enhance data discoverability.
- All aspects of the evaluation and the recommendations made in this thesis.

I was the project manager for this study and was also responsible for:

- defining and describing the project work packages descriptions and deliverables,
- writing the ‘Six criteria for assessing data discoverability’ and ‘Key findings from qualitative analysis of free-text responses’ information boxes for both the reports
- writing the following appendices for the final report: a) data documentation and access: characterising current practice, and b) profiles – documenting cohorts and data resources.
- helping to raise awareness of the study and presented the project at the Public Health Research Data Forum meeting at the Wellcome Trust in January 2014.
- providing feedback, and suggestions for change and improvement to the final report. Drafting of the final report involved the entire project team.

1.6.2 Metadata quality study

I was responsible for all aspects of this study.

1.6.3 Consent for record linkage metadata study

I was responsible for all aspects of this study.

1.7 Ph.D. project funding

The Ph.D. project was funded by a 4-year Medical Research Council CASE award with AIMES Grid Services CIC and partially by the UCL Institute of Health Informatics. The enhancing data discoverability study was competitively commissioned by the Wellcome Trust on behalf of the Public Health Research Data Forum.

1.8 Ph.D. project timeline

The following outlines when each research case study began and finished:

- The consent for record linkage metadata study began in January 2012 and was completed in September 2015.
- The data discoverability study began in January 2014 and chapter in this thesis in September 2015.
- The metadata quality study began in July 2014 and finished in September 2015.

1.9 Chapter summaries

Chapter 2 examines the application of information standards in epidemiological and public health research There are four components to this examination: a) health information standards including encoding and exchange standards; b) linked data on the World Wide Web and the role of Semantic Web technologies in epidemiological and public health research; c) clinical conceptual modelling; and d) application of metadata standards within epidemiological and public health research settings.

Chapter 3 presents research case study 1: enhancing the discoverability of epidemiological and public health research data. This work has four components: a) a systematic review of existing approaches to data discovery and a survey aimed at stakeholders in public health and epidemiology research to identify current data discoverability practices and their subsequent implications; b) qualitative analysis of data in terms of the awareness of the challenges associated with data discovery and identify information; c) presentation of models to enhance research data discoverability; and d) evaluation of the models.

Chapter 4 presents research study 2: improving metadata quality assessment in public health and epidemiological research. This study has four components: a) a systematic literature review of existing approaches to metadata quality evaluation and a survey aimed at stakeholders in public health and epidemiology research in order to identify current practice and challenges associated with creating metadata; b) identification of metadata quality dimensions; c) creation of novel models and framework for assessing metadata quality in epidemiological and public health research settings; and d) evaluation of the novel framework.

Chapter 5 presents research case study 3: improving the recording of consent to record linkage metadata in longitudinal studies. There are four components to this study: a) a systematic literature review of developing models to capture consent for record linkage in longitudinal studies; b) qualitative assessment of consent forms; c) critical evaluation of DDI 3.2; and d) creation and evaluation of the model through iterative application to three test cases.

Chapter 6 presents a summary of my findings and recommendations for change in epidemiological and public health research data management policy and practice. This chapter then presents the overall Ph.D. project strengths and weaknesses followed by a discussion of future direction.

Appendix - there are four appendices to this thesis: a) supplementary tables and a copy of the data discoverability survey; b) a copy of the metadata quality survey and supplementary tables; c) supplementary tables and supplementary code for the consent for record linkage metadata study; and d) a grant proposal for the development of a global metadata registry for epidemiological and public health datasets derived from observational studies.

Chapter 2 Information standards in epidemiological and public health research

2.1 Introduction

In the previous chapter, I introduced the thesis. In this chapter, I examine application of health information standards in epidemiological and public health research focusing on the use of encoding and exchange standards. I then look at the use of linked data on the World Wide Web, clinical conceptual modelling and metadata standards within epidemiological and public health research.

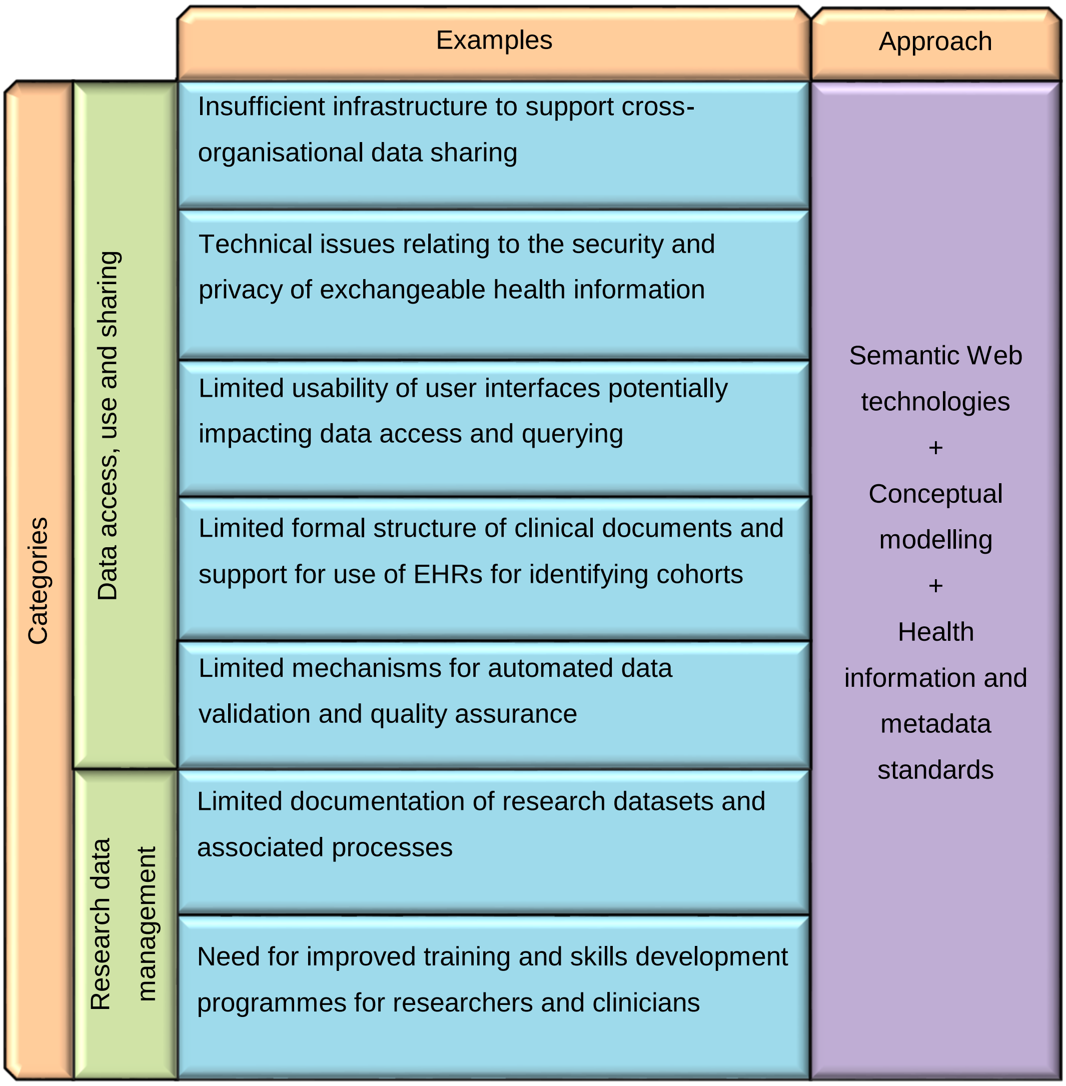
2.2 Standardisation

Standardisation in the public health and epidemiology research domains plays a key role in helping to form platforms or benchmarks from which quality and other such outcomes may be measured and compared (Swensen, Meyer et al. 2010). Impetus to develop and implement health information standards stems from the need to address problems associated with accelerating translation of research findings to policy recommendations, ensuring quality of care and clinical information(Singh, Singh et al. 2013) and implementing a robust information infrastructure to support clinical research (Richesson and Krischer 2007). The use of standards and standardised tools was identified as a key priority in a report by the World Health Organization (2007) and recognised as fundamental to the provision of interoperable health information by the European Commission (2013).

2.3 Challenges associated with using information standards in epidemiological and public health research

Within the UK and internationally, barriers to the adoption of information standards in epidemiological and public health research include: a) ensuring changes to the working environment are appropriately managed on both a technical and staff level (Timimi, Falzon et al. 2012; Singh, Singh et al. 2013); b) having the necessary resources to resolve issues relating to the security and privacy of exchangeable health information; c) having sufficient funding and contingency plans to address any potential financial constraints; d) provision of adequate training and guidance; e) access to the required resources for data documentation and data sharing; and f) limited usability of user interfaces(Heidorn 2008; Rahmouni, Solomonides et al. 2010; Blumenthal 2011; Simborg, Detmer et al. 2013). Figure 2-1 provides a synopsis of the challenges associated with the use of information standards in epidemiological and public health research.

Figure 2-1 Challenges associated with use of information standards in epidemiological and public health research



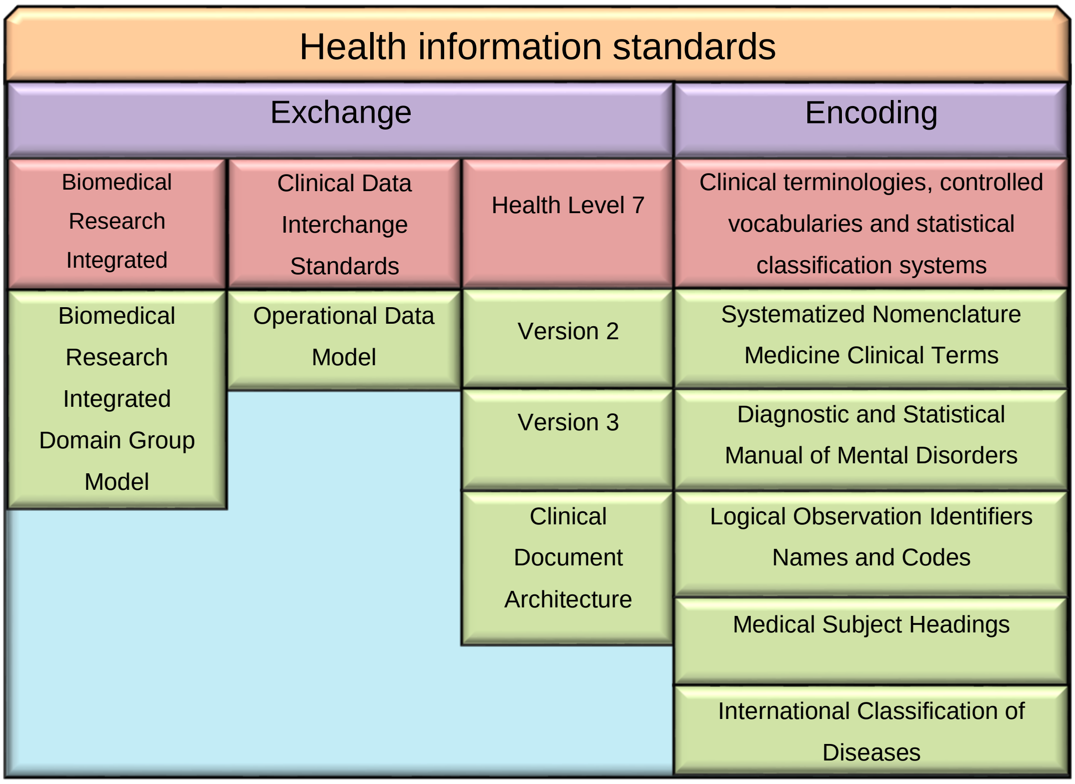
Categories	Examples	Approach
Data access, use and sharing	Insufficient infrastructure to support cross-organisational data sharing	Semantic Web technologies + Conceptual modelling + Health information and metadata standards
	Technical issues relating to the security and privacy of exchangeable health information	
	Limited usability of user interfaces potentially impacting data access and querying	
	Limited formal structure of clinical documents and support for use of EHRs for identifying cohorts	
	Limited mechanisms for automated data validation and quality assurance	
Research data management	Limited documentation of research datasets and associated processes	
	Need for improved training and skills development programmes for researchers and clinicians	

Exemplar challenges associated with use of standards in research and potential approaches to address these.

2.4 Health information standards

There are two broad categories of health information standards: a) encoding – standardises free text found in clinical documents using controlled clinical terminologies which are a type of controlled vocabulary utilised to systematically organise information and support knowledge management); and b) exchange - provide the semantics needed to exchange clinical information/messages (Martin 2003; Tenenbaum, Sansone et al. 2013). Figure 2-2 provides examples of exchange and encoding standards:

Figure 2-2 Exchange and encoding standards



Health information standards

Exchange

Encoding

Biomedical Research Integrated

Clinical Data Interchange Standards

Health Level 7

Clinical terminologies, controlled vocabularies and statistical classification systems

Biomedical Research Integrated Domain Group Model

Operational Data Model

Version 2

Systematized Nomenclature Medicine Clinical Terms

Version 3

Diagnostic and Statistical Manual of Mental Disorders

Clinical Document Architecture

Logical Observation Identifiers Names and Codes

Medical Subject Headings

International Classification of Diseases

2.4.1 Encoding standards

Controlled clinical terminologies such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) are a type of encoding standard. SNOMED CT is primarily used in clinical care to encode clinical information in EHRs. SNOMED CT has a hierarchical structure containing over 311000 concepts. In using a controlled clinical terminology such as SNOMED CT, the way in which clinical information is encoded in EHRs is standardised. Within the research context, this is advantageous as encoded, structured text can be entered into databases, queried and analysed. Additionally, if researchers are presented with free text, by encoding the information using a controlled clinical terminology like SNOMED CT, the now encoded text can become an additional source of information.

Medical Subject Headings (MeSH) is a controlled vocabulary system used to categorise biomedical concepts to support the indexing of biomedical literature from 5400 journals. MeSH has 16 descriptor categories alphabetically organised in a hierarchy.

However, managing controlled clinical terminologies can be problematic. Issues relating to consistency (World Health Organization 2014b) and adequate provision can negatively affect the quality of the encoded clinical information. Another problem associated with the use of clinical terminologies is achieving a consensus with regards to their content. Recent revisions made to DSM 5 included removal of the bereavement exclusion; consequently, a patient suffering with depression, in accordance with the structure and definitions of DSM 5, could be coded as having Major Depression – something certain stakeholders in mental health research and clinical practice stated was not necessarily appropriate (Nemeroff, Weinberger et al. 2013).

Another example of encoding standards are statistical classification systems such as the International Classification of Disease (ICD). ICD is used to capture and classify diseases for clinical epidemiological and management reasons as part of health records and death certificates and billing. The diseases are classified alphabetically and have a hierarchical structure. The Hospital Episode Statistics dataset, which can be used to investigate patterns in disease and the delivery of care in NHS hospitals in England, uses ICD-10 to classify diseases and OPCS-4 ontology to classify procedure codes. The provision of encoded clinical information on a national scale enables researchers to perform in-depth analyses using a rich set of data which have been structured through the controlled formation and expression of clinical information. For example, Raine, Wong et al. (2010) were able to investigate the extent to which hospital admission and surgical procedures varied by factor such as socioeconomic circumstances, age and type of cancer. The authors found that social factors strongly impacted the accessing of care even though the NHS Cancer Plan had been implemented.

Classification systems are complex and subject to change and yet to make meaningful comparisons requires consistent encoding. For example, Cimino (2011) uses the example of septic shock being encoded as 785.52 in ICD-9-CM after 2003; whilst, before 2003 this condition was encoded as 785.59 - other shock without mention of trauma. Individuals wanting to identify cases of shock would need to ensure all possible codes from each of the revisions are included in the queries to help capture all cases. In circumstances such as these, semantic metadata play a key role in helping to integrate heterogeneous data. Integrating multiple datasets with disparate use of these terminologies is commonplace in public health and epidemiological research necessitating the provision and use of crosswalks and other such semantic mappings to enable this process.

A systematic literature review by Stanfill, Williams et al. (2010) on automated clinical coding and classification systems found that the context and complexity of the coding are important factors when evaluating these systems. Of the 113 studies selected for review, examples of clinical terminologies identified included, MeSH terms. The four most commonly identified systems formed 91% of the named systems (46 systems named, 21 not named) with Pneumonia being the most common use case. Studies such as these highlight the importance of encoding standards in clinical practice and research and how their active development and management play a key role in attaining widely accepted, clinically valuable resources.

Table 2-1 presents a comparison of five encoding standards. These are: International Classification of Diseases (World Health Organization 2013a), Systematized Nomenclature of Medicine Clinical Terms(International Health Terminology Standards Development Organisation 2013), Medical Subject Headings(United States National Library of Medicine 2013a), Logical Observation Identifiers Names and Codes(LOINC 2013), and Diagnostic and Statistical Manual of Mental Disorders 5(American Psychiatric Association 2014).

Table 2-1 Comparison of encoding standards

Characteristic	ICD	SNOMED CT	MeSH	LOINC	DSM
Responsible organisation	World Health Organization	SNOMED CT is owned and maintained by the International Health Terminology Standards Development Organisation (IHTSDO)	United States National Library of Medicine (NLM)	Regenstrief Institute (1994)	American Psychiatric Association
Use	Capture and classify diseases for clinical, epidemiological and management reasons as part of health records and death certificates and billing	Capture clinical information in a standardised way to support healthcare provision	Catalogue literature from 5400 biomedical journals as part of the MEDLINE/ PubMED database (U. S. National Library of Medicine 2015)	Capture clinical and laboratory observations as used by >27000 users in 158 countries (2013b)	Capture and classify mental health conditions
Versioning*	Current version - 10 Yearly updates made with version 11 currently underway(2015c). Other classifications include ICD-9-CM	Current version - January 2014 Updates for the International Release are available in January and July every year(2014g)	Current version – 2014 The MeSH browser is updated every Sunday(2014e)	Current version - 2.46 New versions are released in June and December each year(2014d)	Current version – 5 The previous version (DSM-IV) was released in 1994(2014c)

Characteristic	**ICD**	**SNOMED CT**	**MeSH**	**LOINC**	**DSM**
Size	14,199* codes (inclusive of Chapter XX (World Health Organization 2014a)	>311,000* concepts (International Health Terminology Standards Development Organisation)	27,149* descriptors (U. S. National Library of Medicine 2015)	73,115* terms (Vreeman 2013)	20 disorder chapters
Structure	Classified by disease/alphabetical and hierarchical	Hierarchical structure with underlying concept model (International Health Terminology Standards Development Organisation 2014)	16 descriptor categories, alphabetical and hierarchical (United States National Library of Medicine 2013b; U. S. National Library of Medicine 2015)	Each code has 3-7 characters. LOINC names have 6 parts: Component, Property, Time, System, Scale and Method	Each DSM code has a corresponding ICD code
Example	I21.2 Acute transmural myocardial infarction of inferior wall	174041007 laparoscopic emergency appendectomy	Endocarditis, Non-Infective Tree number: C14.280.282.703	67293-1 Other MRI scan [PhenX] Component: Other MRI scan Property: Type Time: Pt System: ^Patient Scale: Nom Method: PhenX	Language disorder 315.32 (F80.2)

* Correct March 2014.

2.4.1.1 Exemplar applications of encoding standards in epidemiological and public health research

Clinical cohort phenotyping: A recent review looking at approaches to identifying patient cohorts by Shivade, Raghavan et al. (2013) found that whilst use of standardised clinical terminologies were promoted in the identified studies, only a few had used them. The authors also identified a lack of metadata describing how certain concepts and terminologies had been mapped together. Subsequently, researchers wanting to identify these descriptions would need to conduct further investigation. This could potentially be a challenging process as due to the lack of metadata, researchers may not have to hand the necessary information to conduct these further investigations and subsequently make efficient and effective use of the data. Having access to good quality semantic metadata could enhance clinicians' and researcher' understanding of the methods used to link the concepts and terminologies and potentially provide them with contextual information which could impact use of the mapped resources as part of epidemiological and public health research studies.

Genomics: Initiatives such as the eMERGE (Electronic Medical Records and Genomics) network promote use of EHRs (Pathak, Wang et al. 2011; Pathak, Kiefer et al. 2012b) and examines the mapping of data elements to standardised metadata repositories and clinical terminologies. However, automated processing of clinical information is not without difficulty. Abbreviations, ambiguity and misspellings limit the extent to which aggregate information may be produced, and subsequently used in clinical practice and research such as identifying clinical phenotyped cohorts. This is because having any abbreviations, ambiguous descriptions or misspelling can cause researchers to inadvertently exclude the records of participants which could potentially be included in the cohort. The result of this is an under-identified cohort of participants with a reduced set of clinical data that could be repurposed and used as part of a research study. Furthermore, having incomplete and erroneous datasets can impact the extent to which metadata standards may be utilised to create metadata.

Epidemiology: The CALIBER (Denaxas, George et al. 2012) resource comprises of linked bespoke studies and electronic health records providing researchers with an in-depth source of information to investigate disease in UK populations. CALIBER contains clinical information from CPRD (2016f), the Myocardial Ischemia National Audit Project (MINAP) (Herrett, Smeeth et al. 2010), Hospital Episode Statistics (HES) (2016i), and the Office for National Statistics (ONS) mortality and social deprivation data. By combining these disparate datasets, researchers are better able to maximise the research benefits associated with record linkage to investigate cardiovascular disease. However, potential inconsistencies in the way in which clinical information was encoded in the primary sources (CPRD, MINAP, HES and ONS) could cause a decrease in data quality. Using data of less than optimal quality can lessen the extent to which research findings can inform cardiovascular policy and clinical practice. Furthermore, as the data is gathered from disparate sources, creating metadata becomes a challenging process. This is because the data originates from different systems in different locations where approaches to data management vary.

2.4.2 Health information exchange standards

A number of organisations focus on developing and maintaining standards in support of exchanging clinical information (Table 2-1). Health Level 7(2015b) is responsible for the development and maintenance of standards in support of the exchange of clinical data. Accredited by the American National Standards Institute, HL7 version 2 is characterised by the use of ASCII code to separate lines of text and is designed to capture patient information. Its uptake on an international level has been very successful (Kalra 2006; Al-Enazi and El-Masri 2013). However, a lack of interoperability and implementation of version 2 led to the development and implementation of version 3 (Kalra 2006).

Messages conforming to version 3 are marked up using XML (Al-Enazi and El-Masri 2013). All developmental work using version 3 is built on or around the BS ISO/HL7 27931:2006 Health Informatics HL7 version 3 - Reference Information Model (RIM) (British Standards Institution 2007). The

RIM is an object oriented abstract model used to derive suitable models to represent healthcare information (Al-Enazi and El-Masri 2013). The conceptual model forms the basis of all version 3.0 derived models and any subsequent development work. The RIM is a platform independent, logical representation of the health information domain facilitating exchange and management. Its structure is static and promotes interoperability between systems built on or around the model (British Standards Institution 2007; Smith, Ashburner et al. 2007; Health Level Seven International 2013). However, whilst HL7 promotes interoperability in clinical settings, the inability to fully characterise diseases or proteins, as discussed by Al-Enazi and El-Masri (2013), Smith, Ashburner et al. (2007) and Smith and Ceusters (2006) limit its application. This is because there is the potential for the minutia needed to perform detailed, complex investigations using messages marked up in version 3 to be missing. This then potentially limits the scope for these messages to be used as additional clinical data sources for epidemiological and public health research studies.

Other models managed by HL7 include the Clinical Document Architecture (CDA) (HL7 2014) and Fast Health Interoperable Resources (FHIR) (HL7 2013). The CDA(HL7 2014) is a mechanism for exchanging healthcare information. By utilising Sematic Web technologies, version 3 RIM and clinical coding schemes, exchangeable documentation has the semantics for machine-readability and clarity for researchers and clinicians. The FHIR framework furthers the progress made as part of the development and maintenance of versions 2, 3 and the CDA to provide a modular approach to building health informatics products (HL7 2013). FHIR provides more flexibility in the way in which health informatics products are designed and built. Consequently, there is the potential for researchers to build products specifically designed to meet their needs.

The Clinical Data Interchange Standards Consortium (CDISC) is responsible for developing and maintaining a collection of standards to support the exchange of clinical data (CDISC 2013b). The Operational Data Model (ODM) is a conceptual model acting as a mechanism for the exchange and archival of data and metadata across disparate sources (CDISC 2013c).

The Biomedical Research Integrated Domain Group (BRIDG) (BRIDG 2012a) is responsible for harmonisation work between certain current health information standards facilitating the development of interoperable applications based on a harmonised conceptual model (Richesson and Krischer 2007; Kush, Helton et al. 2008; McCay, Evans et al. 2008). This work is a collaborative initiative involving stakeholders from CDISC, HL7 Clinical Research Information Management Work Group, National Cancer Institute and United States Food and Drug Administration.(BRIDG 2012b) Drawn using Unified Modelling Language and informed through user feedback, it is a platform-independent model which can be implemented locally to meet specialised user requirements. Furthermore, the provision of a Web Ontology Language (OWL) representation of the model (CDISC 2013a) usability and interoperability are potentially increased through the provision of a machine-readable format. Initiatives such as these bring together expertise across the life sciences domain to promote common understanding. Development of the BRIDG model influenced design of the Life Sciences Domain Analysis Model and the way in which this model captures information (BRIDG 2012c; Freimuth, Freund et al. 2012).

2.4.2.1 Impact of exchange protocols in epidemiological and public health research

The use of electronic systems provides opportunities for automated data validation as a quality assurance mechanism, improving the management of clinical documents and enhancing research (Timimi, Falzon et al. 2012; Ohno-Machado 2013). In 2012, the World Health Organization published a guideline on the (re)development and use of electronic recording and reporting systems for Tuberculosis (TB) care and control (World Health Organization 2012) and in a study by Timimi, Falzon et al. (2012), the potential differences in ways to implement systems and the comparability of clinical data are presented. The current disparate use of electronic systems emphasises the need for standards; yet, a number of challenges associated with their implementation remain (Simborg, Detmer et al. 2013). Tenenbaum, Sansone et al. (2013) discuss the pre-assessment of standards and offer potential categories of criteria such as adoption and user community, in addition to potential resources such as quality assurance tools.

Initiatives such as BioSharing(2014b) map certain life sciences standards together and monitor the management, implementation and reference to these within policies. This is in support of enhanced discoverability and fewer instances of redundant endeavours (2014b). BioSharing has three registries: a) policies; b) standards, from BioPortal, MIBBI and Equator Network; and c) databases in BioDBCore records (BioSharing 2014b; BioSharing 2014c; BioSharing 2014a).

Standards which support communication include ISO 13606(BS EN ISO 2012) and BS EN ISO 13120:2013 (The British Standards Institution 2013). ISO 13606 - Electronic health record communication(BS EN ISO 2012) has been designed to support exchanges of clinical information using the HL7 version 3 standard. BS EN ISO 13120:2013 Syntax to represent the content of healthcare classification systems standard (The British Standards Institution 2013) focuses specifically on standardising the organisation of classification systems and their application to clinical coding schemes such as ICD.

Good quality descriptions (metadata) of clinical terminologies, modelling, data collection methods and the mechanisms are needed to help ensure the security and privacy of participant/patient identifiable data. The development and use of a domain specific metadata quality assurance framework inclusive of Semantic Web technologies, conceptual modelling and health information standards could help researchers and clinicians to create such descriptions.

2.5 Linked data on the World Wide Web

Clinical documents often contain a combination of structured and free-text fields. Consequently, clinicians are able to record both standardised and unstandardised clinical information. To maximise the potential research opportunities associated with these clinical documents, clearly defined underlying structures are needed. Increasingly, Semantic Web technologies are being utilised to help provide these structures to enable researchers to more easily use these as sources of data in epidemiological and public health research (Schweiger, Hoelzer et al. 2002).

Sematic Web Technologies (SWTs) are a group of methods that support the formation of semantically meaningful associations and relationships. Initiatives such as the World Wide Web Consortium Semantic Web Health Care and Life Sciences Interest Group (HCLS IG) (W3C 2013) develop, maintain and promote use of SWTs in the life sciences, healthcare and research. The HCLS IG highlight the important role SWTs play in translational medicine and data linkage (Semantic Web Health Care and Life Sciences Interest Group 2011).

Two examples of SWTs are the Resources Description Framework(2014f) (RDF) and Web Ontology Language(W3C 2012) (OWL). The RDF was developed by the W3C and has been a recommended standard since 2004 for the representing and linking together of heterogeneous data (2014f). The RDF is based on a simple model and has formally constructed semantics. RDF triples may be connected together to form a network. This network can be extended and new information added without negatively impacting the pre-existing structure. The RDF is particularly suitable for metadata management in health as its scalability can be harnessed facilitating the integration of multiple disparate resources. Health data can be stored in databases and spreadsheets across multiple organisations; the application of RDF to create information networks about these data can better support users in discovering health data and exploring its reuse potential particularly for research (Wang, Gorlitsky et al. 2005; Pathak, Kiefer et al. 2012a; Pathak, Kiefer et al. 2012c).The OWL(W3C 2012) is a declarable language enabling clinicians and researchers to develop domain-specific ontologies to assist the capture and structuring of clinical information (Fernandez-Breis, Maldonado et al. 2013).

2.5.1 Role of Semantic Web technologies in epidemiological and public health research

SWTs such as the RDF and OWL provide the mechanisms needed to build searchable, structured networks of information online. SWTs have since been identified as fundamental to the development of scalable solutions to problems associated with the integration of pharmacogenomics, drugs and other such biomedical data (Samwald, Coulet et al. 2012). However, there has not yet been widespread uptake of SWTs in epidemiological and public health research. This outcome could possibly be attributed to the complexity of using SWTs in epidemiological and public health research and the need for additional resources such as time, finance and staff training in order for researchers to benefit from associated research opportunities.

A study by Pathak, Kiefer et al. (2012b) investigated the role of SWTs in the identification of phenotypic data to enable analysis of genetic associations. This work involved a combination of converting clinical information into RDF and using SPARQL to query the information. This initiative utilised the HCLS IG's Translational Medicine Ontology (TMO) (Luciano, Andersson et al. 2011) and other such ontologies to identify patients diagnosed with Type 2 Diabetes Mellitus. Here the authors were able to replicate the findings from a study by Warodomwichit, Arnett et al. (2009) and in this context, acts as a proof of concept that SWTs can assist the identification of cohorts.

Application of SWTs in public health and epidemiological research remains a challenging process. Free-text found in clinical documents are key sources of information; however, providing consistent and high quality annotation is difficult if the information is unstandardised and manually processed. A potential solution is to provide semantic annotations. Tools such as Semantator provide the methods needed link the information to concepts in ontologies to enable processing (Tao, Song et al. 2013). Another tool is the Open Biomedical Annotator (Jonquet, Shah et al. 2009) which focuses on annotating metadata with ontological concepts which are subsequently displayed to users as annotations. Tools such as these can provide a formalised manner in which annotations are generated and subsequently applied. Having automated methods such as these can potentially reduce the risk of human error and biased markup. Nevertheless, use of these tools necessitates additional training and access to resources to support their continued implementation and use.

Awareness of SWTs is increasing and use of these in the biomedical context is important if the potential benefits of standardisation and interoperability are to be realised across the research data lifecycle (Post, Roos et al. 2007; Sagotsky, Zhang et al. 2008; Semantic Web Health Care and Life Sciences Interest Group 2011; Machado, Rebholz-Schuhmann et al. 2013).

2.6 Clinical conceptual modelling

Harmonising definitions of commonly used health-related terms can prove challenging; particularly if there is a lack of, or access to, contextual information (Freimuth, Freund et al. 2012). Having common terminologies and shared understanding can help stakeholders to maximise potential research opportunities in clinical settings through a standardised and simplified approach to mapping information workflows. One such way to map information workflows is to develop clinical conceptual models.

Clinical conceptual models comprise of a series of components and the relationships between these. Application of formalised modelling techniques in epidemiological and public health research settings can help harmonisation and data linkage efforts by providing a simplified yet highly detailed description of clinical settings and the information flows within them. These promote a shared understanding through application of standards applicable across the research domains (Daniel, Sinaci et al. 2014).

2.6.1 Case studies: conceptual modelling in epidemiological and public health research

The following are examples of where conceptual modelling has played an intrinsic role in epidemiological and public health research. In a study by Vawdrey, Weng et al. (2014) conceptual modelling was used to illustrate how common data elements can integrate with existing documentation around EHRs.

Second, the Clinical Research Informatics (CRI) conceptual model (Kahn and Weng 2012) is designed to provide an insight into developments in clinical research informatics. The CRI conceptual model combines clinical and translational research workflows with informatics principles and methodologies to produce a framework. The model centres on workflows, data sources and platforms and informatics core methods and topics (Kahn and Weng 2012). The model also identifies key topics in CRI with secondary use of clinical data, record linkage and data integration being three such examples.

Third, the Life Sciences Domain Analysis Model (LS DAM) (Freimuth, Freund et al. 2012) is designed to facilitate semantic interoperability between several of the Cancer BioInformatics Grid (caBIG) applications. The model can be decomposed into several key areas: Specimen, Molecular Biology, Experiment and Molecular Databases (Freimuth, Freund et al. 2012).

2.6.2 Clinical information models

Standardised modelling techniques can also be used to help develop clinical information models (CIMs). CIMs provide the structural and semantic details needed to facilitate the documenting of clinical concepts (Moreno-Conde, Moner et al. 2015). Collaborations such as the Clinical Information Modeling Initiative (2015a) focus on supporting and improving semantic interoperability within healthcare systems by developing and using collaborative information models. CIMs can also form the basis of information standards such as Health Level 7 and metadata standards such as ISO/IEC 11179.

2.7 Metadata standards

Metadata can be made available in a range of formats such as Portable Document Format (PDF) and eXtensible Markup Language (XML) and can have varying levels of granularity such as research study, single dataset or sweep of data and at variable level. They can be produced in real-time, or retrospectively to reflect one (or more) iteration(s) of the research data lifecycle. Metadata can be classified as: a) administrative - assists with efficient research data management including storage, access and reuse; b) descriptive - describes resources primarily for archival and reuse purposes; and c) semantic - outlines the relationships between metadata elements helping to define underlying structures (Pollock and Hodgson 2004; Zeng and Qin 2008; Miller 2011).

Stakeholders in epidemiological and public health research can systematically describe different metadata elements and their structure using established metadata standards. There are three types of metadata standard: a) structural standards describe the syntactic and semantic requirements needed to ensure the metadata are machine readable; b) content standards support stakeholders in generating metadata by defining what should and should not be included in the description; and c) format standards describe to encode metadata in support of archival and curation efforts(Elings and Waibel 2007; Miller 2011). These standards can be used individually, or combined to form components of a single, much larger standard.

Over the course of a research study, it is possible for metadata standards to change. The use of application profiles, customised versions of metadata schemas, enables stakeholders to implement localised versions of standards to fulfil the stakeholders’ needs. Though this is acceptable, and scope for systematic interoperability with external systems remains possible, unless these changes are approved by the standardisation body responsible for the original standard, the customised version (application profile) may be valid but not formally recognised (Chute, Ullman-Cullere et al. 2013). Therefore, scope for metadata exchange may be negatively impacted subsequently affecting the extent to which the associated research data may be understood by stakeholders and (re)used.

2.7.1 ISO/IEC 11179

The ISO/IEC 11179 (2004) standard supports data reuse and the provision of clearly defined data through a standardised approach to the data management process. The standard is comprised of six parts: 1) framework, 2) classification, 3) registry metamodel and basic attributes, 4) formulation of data definitions, 5) Naming and identification principles, 6) Registration: outlines the registration process.

In applying the ISO/IEC 11179 standards, researchers can potentially benefit from the facilitated monitoring of data to identify similar or identical names and the ability to implement the standard in heterogeneous environments(ISO/IEC 2004). ISO/IEC 11179 was recently implemented as part of METeOR. METeOR is the Australian Institute of Health and Welfare's online metadata registry and is the national repository for public health metadata standards (Australian Insitute of Health and Welfare 2014).

However, in a study by Papatheodorou, Crichton et al. (2009) looking at clinical data management for translational genomics studies in breast cancer using a metadata approach; the researchers found that whilst data elements have associated data element concepts, the standard does not structure the data elements in a particular order. Subsequently, the authors needed to create additional rules to link multiple common data elements and determine which the inferred common data elements were.

Generic metadata standards such as ISO/IEC 11179 do not possess the mechanisms needed to address the needs of epidemiological and public health metadata since they were not designed and developed for sole use in a sole domain. Extending the standard so as to meet these needs is a potential workaround solution but ideally, standards developed specially for biomedical research could potentially alleviate problems such as those mentioned above.

2.7.2 Dublin Core

The Dublin Core Metadata Initiative is responsible for maintaining the Dublin Core metadata standard. Simple Dublin Core (DC) consists of 15 elements designed to provide a standardised basic description of a resource. Each element has attributes as described in the ISO/IEC 11179 standard (ISO/IEC 2012; DCMI 2015).

In a study by Song, Park et al. (2014) looking at the development of a health information search engine based on metadata and an ontology, the authors used the simple DC elements as part of their metadata schema to document resources. There are three parts to their schema: General metadata, Content classification, and Relation. The authors extended the DC schema by adding additional elements such as ‘target audience’. The use of the Dublin Core Metadata Initiative Type Vocabulary (a type of ontology) enables health information to be categorised according to type. The authors were then able to cross-walk the terms in the ontology vocabulary to the clinical terminology, SNOMED CT producing a list of 1300 potential terms to describe health information. The authors found that when compared to a pre-existing search engine, the newly developed health information search engine returned fewer yet more accurate results. Here, the use DC metadata elements combined with an ontology-based vocabulary helped produce a more accurate method of searching for and retrieving health information published to the web.

However, since DC is designed to produce simplistic, domain independent descriptions, it too fails to fully meet the demands of public health and epidemiological metadata. Additionally, simple DC, (this refers to use of the fifteen elements only) does not have scope to record meta-metadata (Miller 2011). Having access to this kind of information can help individuals monitor changes made the metadata and potentially improve opportunities for data citation by helping stakeholders to better identify the correct instance of metadata and to whom credit should be attributed.

Another potential disadvantage with the application of simple DC is the limited scope to record semantic details (2009). Semantic information is needed to facilitate interoperability and enable preservation of digital resources; however, simple DC is not designed to record this kind of information. Within biomedical research, qualified DC would be better suited to producing detailed descriptions. Qualified DC is an altered version of DC consisting of some or all of the original fifteen elements in addition to other more context specific elements. For example, a qualified DC schema could include all, or some, of the 15 simple elements plus extra, user-defined elements to enable the recording of, variable level information, and in which catalogues the metadata has been indexed. Using a schema with additional elements such as these can help stakeholders in public health and epidemiological research to create lower level metadata. Nevertheless, once the simple DC schema has been changed, the resulting schema is no longer compliant with the standard. Consequently, opportunities for metadata exchange could potentially be reduced as stakeholders may experience a loss of metadata or for the metadata to become distorted due to this reduction in interoperability.

2.7.3 Data Documentation Initiative (DDI)

The DDI is an XML-based metadata standard and was designed and developed primarily to describe social sciences research data. The standard is schema based and currently there are two versions both incorporating DC elements: DDI-Codebook (DDI-2)(DDI Alliance 2015a) and DDI-Lifecycle (DDI-3)(DDI Alliance 2015b). DDI-2 is generally used to markup relatively short and focused studies retrospectively. DDI-3 encourages a more real-time approach to marking up metadata for longer term studies and provides mechanisms for metadata comparison. An instance of DDI-3 can be decomposed into four sections:

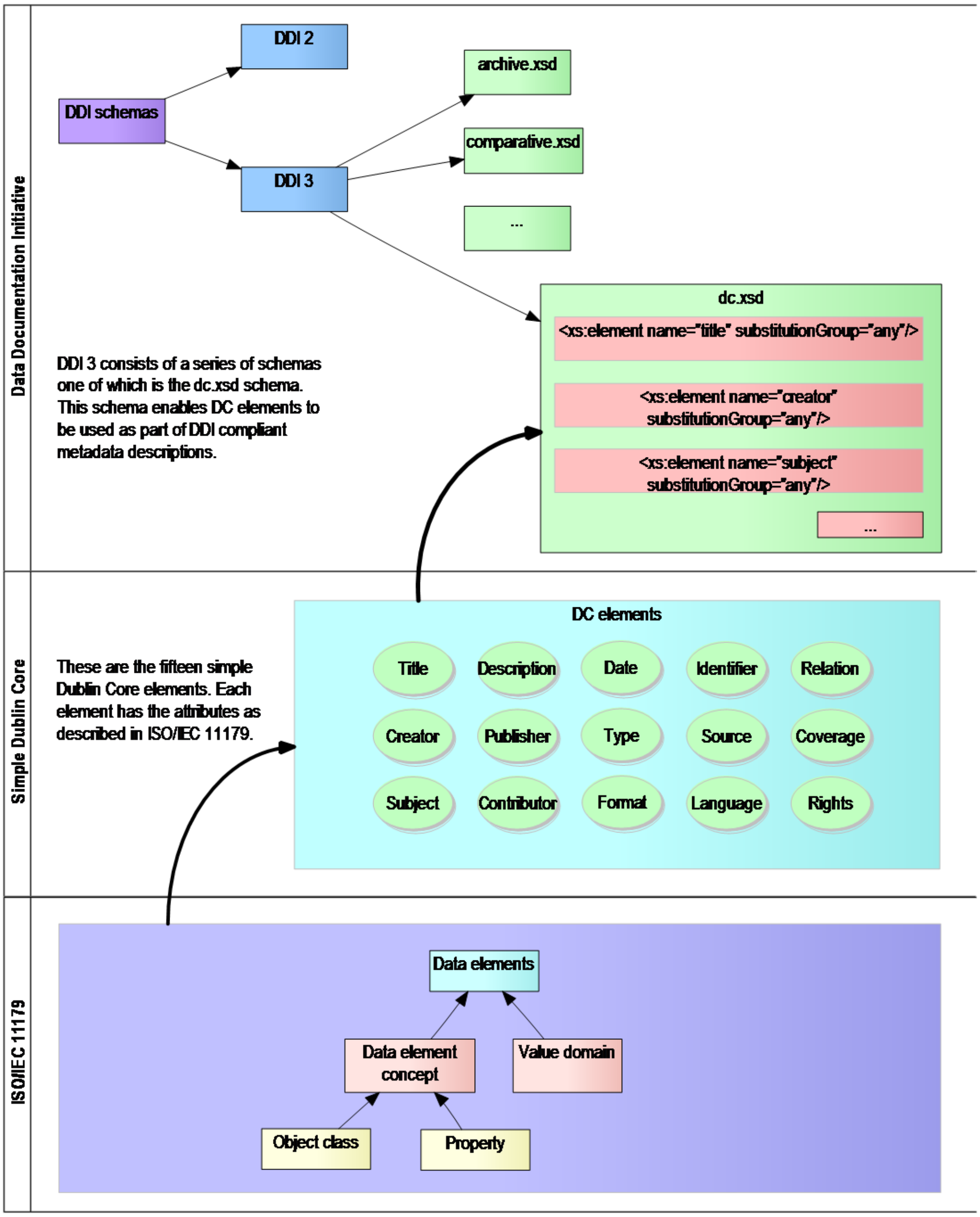
1) Study unit: Comprises of seven subcomponents; conceptual components, inclusive of universes, codes and categories; data collection - describes methodology, question scheme and question logic; logical product, which contains variable information and details of any NCubes (tabulation of variables which is not limited by its own construction); physical data product, physical instance, archive, and profile.(Gregory and Thomas 2012);
2) Group: Holds the resource package and enables maintainable items to be stored collectively. Use of resource packages facilitates inheritance and encourages metadata re-use across different metadata instances;
3) Local holding package: This is the local archive; and
4) Resource Package: Can either be single or multiple published modules containing questions, variables and concepts of a study. For an instance of metadata to be considered valid and compliant, a minimum set of metadata elements must be reached; otherwise the underlying XML will be valid but non-compliant.

The DDI is now commonly used to standardise metadata globally and is the selected standard for international metadata catalogues and archives such as the Consortium of European Social Statistical Data Archives(2016e), Cohort & Longitudinal Studies Enhancement Resources(2016g) and the International Household Survey Network(2016j). The UK Data Archive at the University of Essex also uses DDI to standardise their metadata catalogue records. Application of DDI in this context enabled the creators of these metadata records to provide rich and structured resources. Use of this standard also provides opportunity for metadata harvesting through use of protocols such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

As standards such as the DDI had their origins in the social sciences, they lacked the mechanisms needed to describe epidemiological and public health research studies effectively thus needing modification (Pisani and AbouZahr 2010). By using extensions and customised versions of these standards, this has helped to facilitate their application into epidemiological and public health research settings.

Figure 2-3 shows how these three metadata standards link together.

Figure 2-3 Metadata standards


Data Documentation Initiative
DDI schemas
DDI 2
DDI 3
archive.xsd
comparative.xsd
...
dc.xsd
<xs:element name="title" substitutionGroup="any"/>
<xs:element name="creator" substitutionGroup="any"/>
<xs:element name="subject" substitutionGroup="any"/>
...
DDI 3 consists of a series of schemas one of which is the dc.xsd schema. This schema enables DC elements to be used as part of DDI compliant metadata descriptions.
Simple Dublin Core
These are the fifteen simple Dublin Core elements. Each element has the attributes as described in ISO/IEC 11179.
DC elements
Title
Description
Date
Identifier
Relation
Creator
Publisher
Type
Source
Coverage
Subject
Contributor
Format
Language
Rights
ISO/IEC 11179
Data elements
Data element concept
Value domain
Object class
Property

2.8 Discussion

The repurposing of existing datasets can help to maximise their potential as additional sources of clinical data for use in epidemiological and public health research studies. However, clinical information systems currently lack complete semantic interoperability thus necessitating researchers harmonise disparate data sources before clinical data may be reused in research studies (Bloomrosen and Detmer 2010; Al-Shorbaji 2012; Simborg, Detmer et al. 2013).

Data integration processes are highly complex and demand interoperability on a number of different levels (Anwar and Hunt 2009). One such initiative which addresses these issues, in addition to others such as budget management, is the Clinical Research Administration (CLARA) platform. Bian, Xie et al. (2014) developed the CLARA platform to assist the clinical research management process at the University of Arkansas for Medical Sciences (UAMS) (Bian, Xie et al. 2014). Piloted at the Cancer Institute at UAMS, the CLARA platform now contains 1083 studies of which 91.14% have been fully re-entered into the system.

Further, an increased use of EHRs coupled with the use of incentives, enabled the development of the Query Health initiative (Klann, Buck et al. 2014; Embi, Weir et al. 2013). Query Health aims to address issues relating to performing standardised clinical queries in a secure and environment, implementation and to assess effectiveness through piloting. The pilot studies involved collaborating with the Department of Health in New York City and Massachusetts and the Food and Drug Administration Mini-Sentinel(Platt, Carnahan et al. 2012) program (Klann, Buck et al. 2014). The developers used HL7 and SWTs to allow users to query and the Query Health ontology to manage terms derived from the National Quality Forum’s Quality Data Model (Klann, Buck et al. 2014).

In conclusion, the application of SWTs, conceptual modelling and metadata standards, in addition to health information standards has had an impact on epidemiological and public health research and continues to do so.

2.9 Chapter summary

In this chapter I investigated the use of information standards in epidemiological and public health research. I firstly examined the health information standards and focused on the application of encoding and exchange standards with epidemiological ad public health research settings. I then focused on the use of linked data on the World Wide Web in epidemiological and public health research. I then explored the use of clinical conceptual modelling and the use of clinical information models. Following this, I examined the use of metadata standards, ISO/IEC 11179, Dublin Core and the Data Documentation Initiative in epidemiological and public health settings. In the next chapter, I will describe the enhancing research data discoverability study and present my findings. In this study I investigated ways to enhance the discoverability of epidemiological and public health research data by performing a systematic literature review, online stakeholder survey and identified and evaluated mechanisms to enhance discoverability.

Chapter 3 Research case study 1: Enhancing the discoverability of epidemiological and public health research data

3.1 Introduction

In the previous chapter I examined the role of information standards in epidemiological and public health research. A key theme which emerged from this examination was the reuse of research data and repurposing of clinical data for research purposes. By repurposing and reusing data, stakeholders have the opportunity to maximise the research benefits associated with a cyclical use of research data in epidemiological and public health research settings.

However, current limitations associated with research data discoverability can negatively impact the extent to which certain data are known to the wider research community. Discoverability refers to the ease to which research data may be identified by potential users. Subsequently, users may explore and characterise the data before harnessing or exploiting it potential research opportunities. Consequently, the potential for these data to be utilised for additional research purposes and the potential benefits realised is reduced (MRC 2014).

Case study: ELFE, Growing up in France

The ELFE study focuses on the lives of children growing up in France. It began in April 2011, and has since followed the lives of more than 18000 children over a course of 20 years[1].

Whilst a list of related publications is provided on the study website, detailed metadata could not be found. Furthermore, there are no tools available to search through the metadata and visualise the research data. Consequently, researchers and other stakeholders wanting to discover and explore potentially available research data are unable to do so.

The aim of my research case study is to identify and evaluate mechanisms to enhance the discoverability of public health research data discoverability. These mechanisms may be applied to the studies such as ELFE to enhance their discoverability and the potential to reuse and repurpose research data.

1. http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=326&sitesectiontitle=ELFE+%28Growin%20g+up+in+France

Furthermore, identifying and characterising certain research datasets such as linked clinical data as potential additional data sources, remains a challenging aspect of public health and epidemiology research (Weber, Mandl et al. 2014).

This chapter presents the data discoverability study in which the enhancement of public health and epidemiology data discoverability was examined.

3.2 Study aim and objectives

The aim of this study was to identify and evaluate within the context of public health and epidemiological research settings mechanisms through which research data discoverability may be enhanced. This study had four objectives: a) describe current approaches to making research data discoverable; b) identify current awareness of the data discoverability issue and the perceived challenges of using tools, technologies and catalogues; c) identify models to enhance data discoverability; and d) evaluate these.

3.3 Methods

3.3.1 Systematic literature review

The systematic literature review sought to characterise existing data discoverability practices and to identify current challenges and uses of metadata technologies. I used PubMed, Web of Science, Google and forward citation tracking (Kuper, Nicholson et al. 2006) and the following search terms: ‘public health and epidemiology’, ‘research datasets’, ‘data discoverability’, ‘metadata’ and ‘data reuse’. These terms were selected following a discussion with project team members and background literature. The use of these terms in the survey responses gave a degree of reassurance they were the correct terms to use. Unique identifiers were assigned to each of the studies and organisations as they were identified. A total of 49 (Supplementary Table 1) public health and epidemiological studies and organisations were identified of which 13 were randomly selected and reviewed. Thirteen was a sufficiently large enough number for a range of different studies to be analysed in-depth within the time constraints of the project and to avoid any potential bias. The six point criteria I defined was:

1. Provision of study protocols: To establish the extent to which these were made publically available and in which formats.
2. Approach to data documentation: To determine if these were publically available, in which formats and whether these were available for download
3. Provision of data access mechanisms: To determine how these were made known to the public, identify any access and/or use policies and guidelines, and identify, if any, datasets available for immediate download.
4. Publicly available online data visualisation and/or analysis tools: Firstly to establish whether researchers or other members of the public were able to utilise the data to perform basic analyses. Secondly, the recording of this information could inform development of any subsequent tools to enable researchers to better their understanding of the data before application is made for access and use.
5. Provision of links to or descriptions of, publications: To characterise current approaches to publicising journals papers and other such forms of media to promote data discovery and inform the wider community of research findings.
6. Use of social media and/or other forms of communication: To determine the extent to which social media is used to help improve knowledge of studies and/or organisations and which ones, if any, in particular. This would also inform development of any tools as these act as additional mechanisms through which discoverability may be enhanced.

These criteria were selected as they were most appropriate to the literature review objectives. The use of social media was included as platforms such as YouTube and Facebook are being harnessed (in addition to other mediums) to better communicate with members of the public and (potential) participants of studies. I wanted to investigate use of these as they can contribute to the heightened discoverability of studies thus increasing awareness.

3.3.2 Online stakeholder survey

I have chosen to structure my description according to the CHERRIES(Eysenbach 2004) statement as adopted by the Journal of Medical Internet Research for reporting the results of web-based research as it enabled us to systematically describe in detail the design and administration of the online survey.

3.3.2.1 Design

The survey was composed of six sections informed by findings of the review such as use of metadata standards and data publications. The sections were: a) background information – participant demographics; b) data discoverability – areas of importance, repositories, clinical terminologies and classification systems; c) data repositories; d) controlled vocabularies and thesauri; e) data documentation; and f) data citation and data publications. The survey is self-selecting hence not representative.

In the data discoverability section, the list of repositories given to respondents to select from was based on a list provided by the Nature Publishing Group on their website (Nature Publishing Group 2014). In the data citation and data publications section, the list of perceived benefits was adapted from a report by the Digital Curation Centre (Ball and Duke 2012). A copy of the survey questions can be found in appendix A.

During the pre-testing I found that the Boolean logic needed altering to ensure that certain questions, e.g. 'If other, please specify…' appeared as expected. I also changed the size of the answer boxes to enable respondents to provide more detailed responses. Other aspects of the survey that were adapted following pre-testing included altering the list of forms of data to include forms of data not previously listed. It was assumed that the respondents knew of and understood the term 'metadata' and other related terminology such as 'research data lifecycle'.

3.3.2.2 Ethical approval and informed consent process

Ethical approval was not required for the work undertaken as part of the Ph.D. project. Implied consent to partake in the study was assumed from the individual through their decision to submit data using the online survey. All data collected is anonymous and contact details were provided if (potential) participants wished to contact the project team for further information or clarification.

3.3.2.3 Development and pre-testing

The survey was designed and developed using REDCap version 5.7.5.(Harris, Taylor et al. 2009) REDCap, a web-based data capture tool enabling development of survey instruments and collection of data in a secure environment.

3.3.2.4 Recruitment process and description of the sample having access to the questionnaire

The following mailing lists were used to circulate the survey:

- JISC Research data management
- JISC Managing research data
- JISC Public health mailing list
- JISC UK health and medical library

The survey ran from 31st March 2014 to 21st April 2014. The advantage of using mailing lists is that a large number of people can be contacted in a relatively short period of time. The weakness of this recruitment method is that it is unfeasible to calculate a response rate as the total number of people subscribed to these lists, and to what extent the invitational emails were forwarded was unknown.

In addition to the mailing lists, representatives of the signatories and supporting organisations of the Public Health Research Data forum were also contacted and asked to circulate the surveys. The signatories were: Agency for Healthcare Research and Quality, Bill and Melina Gates Foundation, Canadian Institutes of Health Research, Centres for Disease Control and Prevention, Deutsche Forschungsgemeinschaft (DFG), Doris Duke Charitable foundation, ESRC (UK), Health Research Council of New Zealand, Health Resources and Services Administration (USA), Hewlett Foundation, INSERM, MRC (UK), National Health and Medical Research Council (Australia), National Institutes of Health (USA), South African Medical Research Council, Substance Abuse and Mental Health Services Administration (USA), USAID, Wellcome Trust, and The World Bank. The supporting organisations are: Chief Scientist Office (Scotland), Creative Commons, Emergency Nutrition Network, UNICEF, and World Health Organization.

During an initial stakeholder analysis as part of the initial scoping of the study, four groups were identified: researchers and other data users, data producers, archivists and librarians, and funders. These groups are not mutually exclusive. This analysis also provided us with an opportunity to raise awareness of the survey and to collate additional email addresses of potential survey participants. A total of 113 individuals with a total of 88 different affiliations, were identified and invited to complete the data discoverability survey.

3.3.2.5 Survey administration

The invitational email, inclusive of a brief description of the study and a link to the survey, asked the participant to forward the invitation to their contacts in an attempt to reach as many potential participants as possible.

The survey began with a short introductory paragraph to the study describing the aims and objectives of the survey and the length of time it should take to complete (10 minutes). To facilitate data capture, the survey consists of several pages (Schleyer and Forrest 2000). The questions were grouped according to theme and the participants were able to track how far into the survey they are in terms of pages. All questions were optional and where multiple selections were possible, this is indicated in the question. When completing the survey, the participants had the option of saving and returning to the survey as many times as they deem necessary.

Certain questions invited respondents to provide additional suggestions and comments through use of ‘if yes/other, please specify…’ style questions; Boolean logic was incorporated to automatically customise the survey depending on previously submitted answers, this in turn removed or added any necessary subsequent questions. The inclusion of open-ended questions in the survey facilitated the capture of qualitative data. Furthermore, a Likert scale comprising of five points (‘not at all’, ‘slightly’, fairly’, ‘extremely’, and ‘essential’) was used to gauge stakeholders’ opinions and feedback.

3.3.2.6 Analytical strategy

All completed questionnaires were analysed. I analysed the quantitative data collected through the online survey with SPSS. To analyse the qualitative results, I adopted a Grounded theory approach and the themes were collated inductively and iteratively. This approach enabled us to use to the data collected to develop theories hermeneutically rather than approach the data with hypotheses already formulated.

3.3.3 Mechanisms to enhance discoverability

Based upon the literature review and online stakeholder survey, three models were identified (data publications, linked data and a public health portal) through which funders could enhance data discoverability. Each of these models arose from my research findings. Through my systematic literature review I identified the potential to use data publications as another way of enhancing data discoverability. In this thesis (chapter 3), I identified the potential to use SWTs as mechanisms to further enhance data discoverability. Through qualitative analysis of the survey results, I identified the need for a registration process for observational studies which could be enacted through a public health portal similar to that of ClinicalTrials.gov.

3.3.4 Evaluation strategy

I conducted a series of feasibility analyses (Dennis, Wixom et al. 2015) by way of evaluation. This approach originates from the computer science domain and is used as part of systems analysis and design. In adopting this approach, I was able to systematically and thoroughly analyse three key areas of feasibility to determine how valid each model was in enhancing data discovery. The three areas were: a) technical – factors affecting how the system is built; b) economic – this examines the associated costs/benefits; and c) organisational – factors which could affect how the system is used by stakeholders. In having a clear framework from which to work I was able to identify case studies where these models had been previously utilised in epidemiological and public health research. I also engaged with stakeholders in epidemiology and public health to critically appraise each model by way of evaluation. The mechanisms were presented to members of the Public Health Research Data Forum and The Wellcome Trust through teleconferences and face-to-face discussions.

3.4 Results

3.4.1 Systematic literature review

A total of 49 studies and organisations were identified, of which 13 were randomly selected and their websites reviewed in greater detail. Most of the studies identified are observational, for which there is no known mandatory registration process. Table 3-1 describes the results of the review of the sample of epidemiological and public health studies and organisations.

The review showed that of the sample, PDF was one of the most common formats to provide study protocols. It also showed that SAS, STATA and SPSS were the most commonly supported file formats for the data. Results also show all studies and organisations in the sample provide either a list and/or sometimes link to publications involving particular datasets.

The review showed that five studies and organisations provided online data visualisation/analysis tools: the European Prospective Investigation into Cancer and Nutrition (EPIC), European Social Survey, INDEPTH Network, IPUMS International Project, Measure DHS and the Worldwide Antimalarial Resistance Network. These help prospective researchers and members of the public are able to use certain datasets and perform basic analyses.

Results also show newsletters, RSS feeds, Facebook, Twitter, YouTube and blogs were among the alternative forms of communication employed by the studies and organisations.

Table 3-1 Review results and review according to the 6 point criteria developed

Study ID, name and coverage	Study protocols	Data documentation	Data access	Online data visualisation/ analysis	Publication links/ descriptions	Social media / other forms of communication
6 - Avon Longitudinal Study of Parents and Children UK	Questionnaires available as PDFs	Downloadable data dictionary	Data access policy and guidance available online (PDF)	No	Yes	Facebook, Google+, Soundcloud, MyYahoo, YouTube, Twitter and QR code
12 - ELFE, Growing up in France France	Online description of key stages	Could not find this information on their website	Could not find this information on their website	No	Yes	Newsletter
13 - European Prospective Investigation into Cancer and Nutrition (EPIC) Lyon, UK (London), Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, The Netherlands	Questionnaires and statistical methods	Questionnaires and descriptions of the cohort/anthropometric measurements available online. All measurements were standardised using EPIC-SOFT (available in multiple languages) to enable comparison.	Could not find this information on their website	No	Yes	RSS and LinkedIn

Study ID, name and coverage	Study protocols	Data documentation	Data access	Online data visualisation/ analysis	Publication links/ descriptions	Social media / other forms of communication
14 - European Social Survey Albania, Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Kosovo, Latvia, Lithuania, Luxembourg, The Netherlands, Norway, Poland, Portugal, Romania, Russian federation, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, UK	Online descriptions and PDFs	Available online by year, country and theme	Data is available for download online in SAS, SPSS and STATA	Yes	Yes	Email
18 - INDEPTH Network South Africa, Guinea Bissau, Senegal, Ethiopia, Ghana, Gambia, Tanzania, Uganda, Malawi, Burkina Faso, Kenya, Mozambique, Nigeria, Cote d'Ivoire, India, Bangladesh, Vietnam, Cambodia, Thailand, Indonesia, Papua New Guinea	Study overviews available online	Online data dictionaries available with variable names, labels and descriptions. DDI compliant metadata is available. Downloadable microdata (must register/login)	Two types: 'Public use files' and 'Licensed files' and data can be filtered according to centre.	Yes	Yes	Email

Study ID, name and coverage	**Study protocols**	**Data documentation**	**Data access**	**Online data visualisation/ analysis**	**Publication links/ descriptions**	**Social media / other forms of communication**
20 - IPUMS International Project Argentina, Armenia, Austria, Bangladesh, Belarus, Bolivia, Brazil, Burkina Faso, Cambodia, Cameroon, Canada, Chile, China, Colombia, Costa Rica, Ecuador, Egypt, El Salvador, Fiji, France, Germany, Ghana, Guinea, Greece, Haiti, Hungary, India, Indonesia, Iraq, Iran, Ireland, Israel, Italy, Jamaica, Jordan, Kenya, Kyrgyz Republic, Malawi, Malaysia, Mali, Mexico, Mongolia, Morocco, Nepal, Netherlands, Nicaragua, Pakistan, Palestine, Panama, Peru, Philippines, Portugal, Puerto Rico, Romania, Rwanda, Saint Lucia, Senegal, Sierra Leone, Slovenia, South Africa, South Sudan, Spain, Sudan, Switzerland, Tanzania, Thailand, Turkey, Uganda, United Kingdom, United States, Uruguay, Venezuela, and Vietnam	Questionnaires available as PDFs and HTML	When data are extracted, codebooks are generated.	Online variable selection (integrated and harmonised variables options available). Users must be registered before data extracts may be downloaded. Extract system provides support for the import of generated ASCII files into SPSS, SAS and STATA.	Yes	Yes	Newsletter
22 - Measure DHS, Demographic Health Surveys Angola, Benin, Botswana, Burkina	Questionnaires are available for download (PDF)	Data are recoded (variable names, locations etc.) with all recording	Must be a registered user of the website. Application for	Yes -HIV/AIDS Survey Indicators Database &	Yes	Email, Facebook, Twitter, YouTube,

Study ID, name and coverage	Study protocols	Data documentation	Data access	Online data visualisation/ analysis	Publication links/ descriptions	Social media / other forms of communication
Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Comoros, Congo (Brazzaville), Congo Democratic Republic, Cote d'Ivoire, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guinea, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Nigeria (Ondo State), Rwanda, Sao Tome and Principe, Senegal, Sierra Leone, South Africa, Sudan, Swaziland, Tanzania, Togo, Uganda, Zambia, Zimbabwe, Albania, Armenia, Azerbaijan, Egypt, Jordan, Moldova, Morocco, Tunisia, Turkey, Ukraine, Yemen, Central Asia, Kazakhstan, Kyrgyz Republic, Tajikistan, Turkmenistan, Uzbekistan, Afghanistan, Bangladesh, Cambodia, India, Indonesia, Lao People's Democratic Republic, Maldives, Nepal, Pakistan, Philippines, Sri Lanka, Thailand, Timor-Leste, Vietnam, Samoa, Bolivia, Brazil, Colombia, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Haiti, Honduras, Mexico,		manuals available for download online. Use the DHS Recode.	data must include contact information, research project title and description of intended analysis. Certain data require users to sign additional agreements/terms of use. Files are distributed as compressed Zip files. List of potential data downloads is available online. Data are available in ASCII, STATA, SPSS, SAS	STAT compiler		LinkedIn, Pinterest, Blog and Mobile phone app

Study ID, name and coverage	Study protocols	Data documentation	Data access	Online data visualisation/ analysis	Publication links/ descriptions	Social media / other forms of communication
Nicaragua, Paraguay, Peru, Trinidad and Tobago			and CSPro			
23 - MIDUS Midlife in the US United States of America	Available in catalogue/ICPSR website	Categories, codes, variable grouping etc. available through Colectica catalogue. DDI Codebook/Lifecycle, Dublin Core and MARC21 XML metadata	Data is available for download online in SAS, SPSS, STATA, ASCII and Delimited from ICPSR website	No	Yes	LinkedIn, Facebook, Google or MyData
26 - Norwegian Mother and Child Cohort Study Norway	Online description and downloadable questionnaires	Basic participant response figures available online for version VIII	Applications must be submitted and approved before use of data and/or biological materials is enabled. Researcher(s) may need to sign a contract with the Norwegian Institute of	No	Yes	Facebook, Twitter, YouTube, RSS feed and newsletter

Study ID, name and coverage	Study protocols	Data documentation	Data access	Online data visualisation/ analysis	Publication links/ descriptions	Social media / other forms of communication
			Public Health			
33 - Scottish Longitudinal Study Scotland	Online descriptions of creation/development available as individual 'Technical Working Paper'	Online data dictionary with searchable lists of tables and variables	2 methods - 'safe setting' in Edinburgh or remote access involving use of variable names and labels only for creation of syntaxes. These syntaxes are then returned to the Support Officer who will arrange for the analysis to be run and the results, once cleared returned. More details available online	No	Yes	Twitter, email, blog

Study ID, name and coverage	**Study protocols**	**Data documentation**	**Data access**	**Online data visualisation/ analysis**	**Publication links/ descriptions**	**Social media / other forms of communication**
34 - Study of Environment on Aboriginal Resilience and Child Health Australia	Study protocol available as an academic paper	Could not find this information on their website	Describes 'SURE' Secure, Unified Research Environment' but does not make it clear if SEARCH data can be accessed	No	Yes	Email, Facebook, RSS, LinkedIn and Twitter
48 - WHO Study on Global AGEing and Adult Health (SAGE) South Africa, China, Ghana, India, Mexico, Russian Federation	Summary of measures in questionnaires and the questionnaires themselves available in PDF	Related materials, study description, data dictionary (variables include name, label and question) and related citations all available. DDI compliant metadata available in PDF.	Through WHO Multi-Country Studies Data Archive.	No	Yes	RSS, YouTube, twitter, Facebook and Google+

Study ID, name and coverage	Study protocols	Data documentation	Data access	Online data visualisation/ analysis	Publication links/ descriptions	Social media / other forms of communication
49 - Worldwide Antimalarial Resistance Network Thailand, Kenya, Brazil, Senegal	Searchable procedures available for download	When sharing data (clinical, pharmacology, In vitro) data dictionaries should be accompany data	WWARN standardise data and then make these available including of audit trail and original dataset to the data contributor and nominated individuals. Researchers (under ‘Third party data access’) should contact the data owner(s) as they alone can grant access to the transformed data.	Yes	Yes	Facebook and twitter

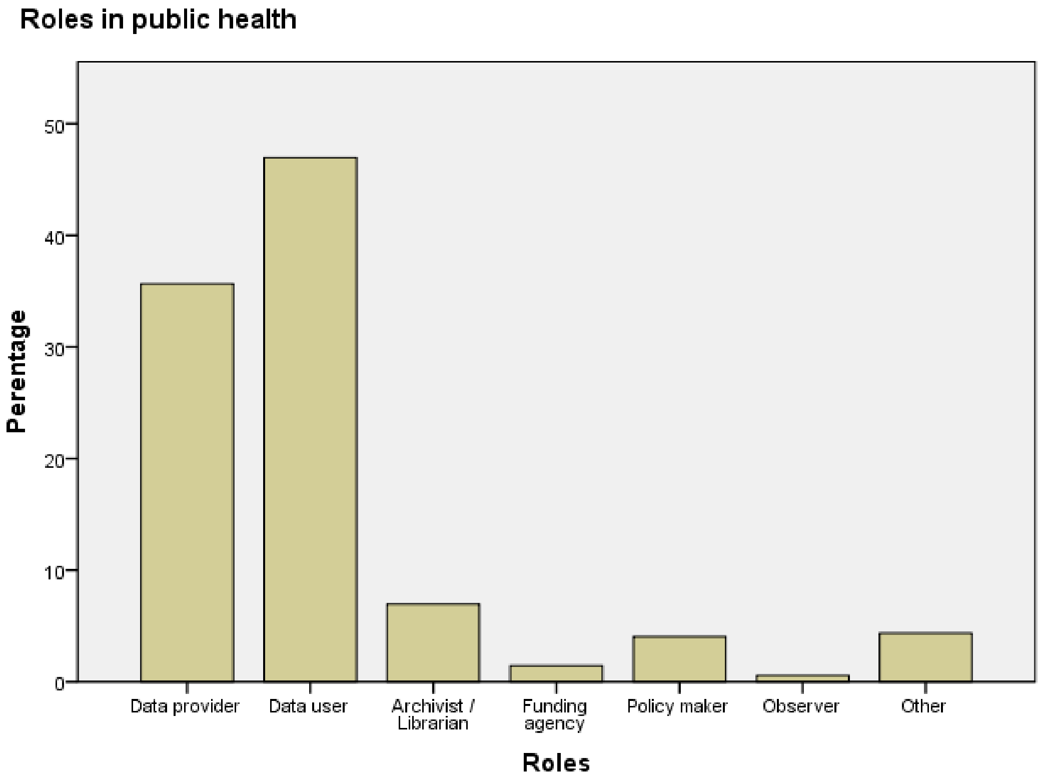
3.4.2 Online stakeholder survey[1]

3.4.2.1 Participant demographics

253 individuals completed the survey of which most were employed by a university, Figure 3-1. The respondents were also asked to indicate their role in public health research data and were able to make multiple selections. The most common role was 'Data user' and the least common was 'Observer'. The most common respondent was a Data User located in Europe. Just over a quarter of survey respondents carried out work in Europe (28.9%) followed by Oceania (13.1%) and Northern America (8.3%). The survey also requested respondents indicate if they were in receipt of funding from any of the specified funding agencies, Table 3-2. The most common funding agency from the pre-determined list was the Medical Research Council (UK) (15%) followed by the Wellcome Trust (11%. A total of 61 respondents indicated that they received funding from other non-specified agencies with the French Ministry of Foreign Affairs, Rio de Janeiro State research Support Foundation (Brazil) and Australian Red Cross Blood Service being three such examples.

[1] N.B. Where quotations have been used, these have been copied verbatim; in places words have been added in parentheses to aid understanding.

Figure 3-1 Roles in public health



Roles are defined as: a) data provider – provides research data; b) data user – someone who uses research data; c) archivist/librarian – anyone who is responsible for cataloguing, curating or storing research data on both a short to medium term and longer term basis; d) policy maker – those involving in policy making and or advising; e) observer – anyone indirectly involved in the research process; and f) other.

Figure 3-2 Location of work undertaken

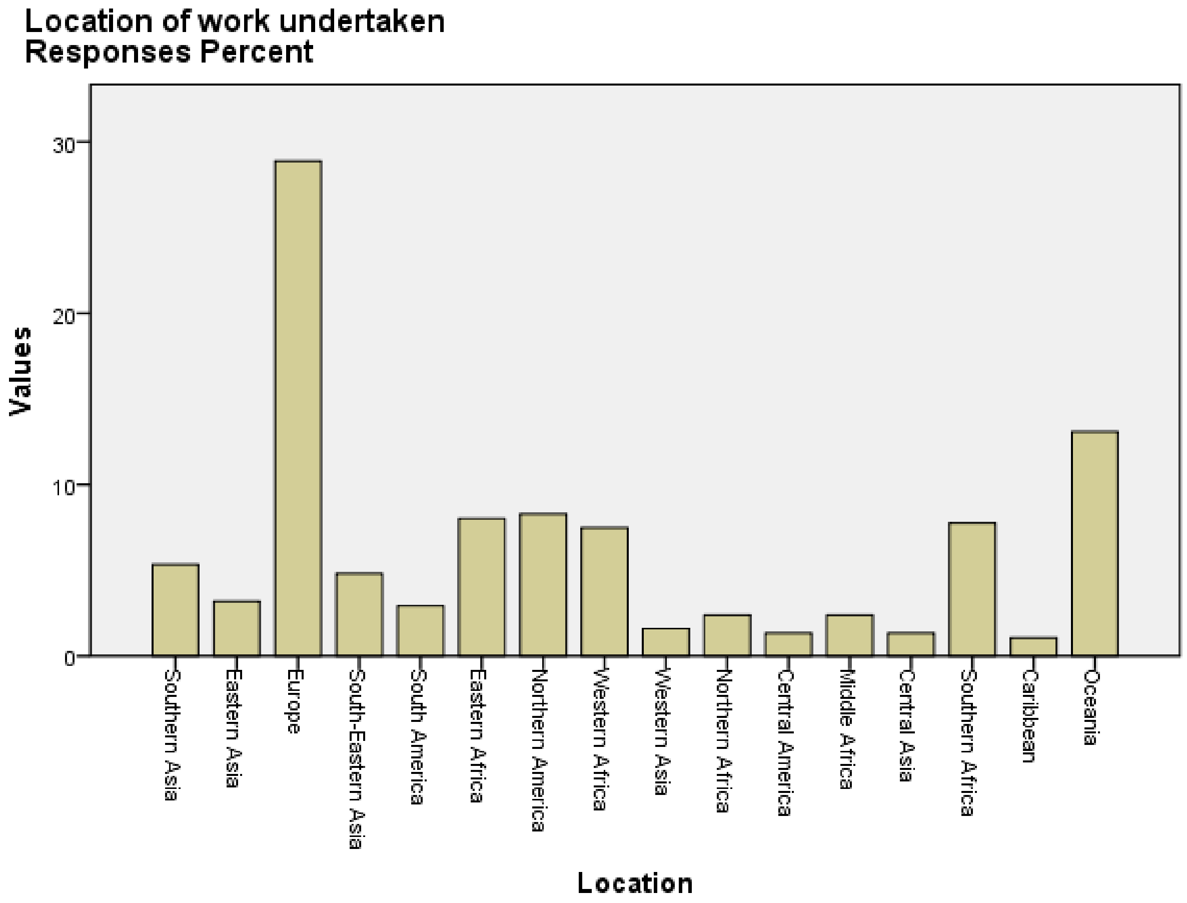


Table 3-2 Funding agencies

	Responses	
	Number	Percent
Medical Research Council (UK)	44	15%
Wellcome Trust	31	11%
National Health and Medical Research Council (Australia)	28	10%
Bill and Melinda Gates Foundation	25	9%
NIHR (UK)	24	8%
National Institutes of Health (USA)	18	6%
Economic and Social Research Council (UK)	18	6%
Centres for Disease Control and Prevention	8	3%
Agency for Healthcare Research and Quality (USA)	4	1%
Canadian Institutes of Health Research	3	1%
Deutsche Forschungsgemeinschaft (DFG)	2	1%
Health Research Council of New Zealand	4	1%
Health Resources and Services Administration (USA)	3	1%
Hewlett Foundation	3	1%
Institut national de la santé et de la recherche médicale (INSERM, France)	2	1%
Substance Abuse and Mental Health Services Administration (USA)	2	1%
The World Bank	4	1%
Other(s)	61	21%

3.4.2.2 Data types and the research data lifecycle

The most commonly used form of data by survey respondents was survey data, (27%), and the second most common were, healthcare records, 21%. The least common was imaging data, 3%. Table 3-3 presents the results (in percent) the different forms of data as used by the respondents who provided data. These results vary however at an individual country level and also by role.

Table 3-3 Forms of data

	Responses	
	N	Percent
Survey	157	27%
Healthcare records	125	21%
Disease registries	76	13%
Ethnographic	24	4%
Geospatial	46	8%
Environmental	31	5%
Genomic/Proteomic/Metabolomic	30	5%
Imaging	19	3%
Physiological measurement	47	8%
Other	37	6%

The respondents were also asked to indicate in which areas of the research data lifecycle were they involved in. Table 3-4 shows the respondents' roles in public health as categorised by location.

Table 3-4 Role in public health and location of work

	Role in public health (number)						
	Data provider	Data user	Archivist / Librarian	Funding agency	Policy maker	Observer	Other
Southern Asia	14	14	1	1	1	0	2
Eastern Asia	8	8	1	1	1	0	1
Europe	64	83	10	2	6	1	6
South-Eastern Asia	9	14	1	1	1	0	2
South America	5	9	1	0	2	0	2
Eastern Africa	22	21	2	3	3	0	2
Northern America	14	16	9	0	1	1	5
Western Africa	17	20	2	3	2	0	1
Western Asia	3	4	1	0	1	0	1
Northern Africa	5	4	1	0	1	0	2
Central America	2	3	1	0	1	0	1
Middle Africa	6	4	1	0	1	0	2
Central Asia	3	2	2	0	1	0	1
Southern Africa	21	19	3	1	1	0	4
Caribbean	2	1	1	0	1	0	1
Oceania	22	42	6	1	5	0	1

3.4.2.3 Search options, controlled vocabularies and thesauri

Another issue explored in the survey was the use of controlled vocabularies and thesauri. This issue was explored because there are inconsistencies between the way in which researchers describe their research and the way in which other stakeholders search.

In certain circumstances, such as describing manuscripts, authors can select MeSH terms as ‘key words’. MeSH terms are used to index medical literature in biomedical databases; however, it is possible for a ‘key word’ associated with a manuscript to not appear in the MeSH term listing. For example, ‘metadata’ does not have an entry in the MeSH term listing[2]. A potential work around here would be to use related subject terms which do appear in MeSH; in the case of ‘metadata’, this could become ‘documentation (L01.453.245)’[3]. Whilst this would not seem problematic, the challenge for stakeholders arises when they want to form search strategies to source literature i.e. how do they know when to use a key word such as ‘metadata’ or a subject term such as ‘documentation’. It is also possible in these circumstances for authors to include ‘concepts’ in an attempt to enhance the description of their paper. For example, a concept could be ‘epidemiology’ and a related concept could be ‘public health’. This does however raise the question of determining when a ‘concept’ ceases to be just that and becomes a ‘related concept’. To the best of my knowledge, the distinction or limits between these has not yet been clearly defined and ratified by the wider scientific community for purposes of describing manuscripts. And yet, some authors need to use these to describe their work as the ‘key words’ needed do not have equivalent ‘subject terms’.

Therefore, should stakeholders treat ‘subject terms’, ‘key words’, ‘concepts’ and ‘related concepts’ as though they are mutually exclusive and develop four different search strategies; or, is it that a single search strategy containing a mix of these will suffice? This does however raise the question

[2] correct January 2016 using 2016 MeSH

[3] correct January 2016 using 2016 MeSH

of at which point should stakeholders sourcing literature draw the boundaries of their search strategy since to use every ‘key word’, ‘subject term’, ‘concept’ and ‘related concept’ (mutually exclusively or otherwise) can become infeasible. Differences in the way authors describe their manuscripts given a lack of clear guidelines could negatively impact the potential for discovery as part of literature searches.

Hence, understanding the thought process behind the way in which authors describe their manuscripts is important to informing recommendations aimed at authors for purposes of enhancing discoverability of their scientific outputs. It is also important to informing recommendations aimed stakeholders in searching for literature for purposes of potential reuse of existing datasets to test additional hypotheses. Respondents were thus asked to indicate which their preferred search options were. For purposes of this study ‘key word’ is defined as a word associated to the manuscript critical to describing its content; ‘subject terms’ are defined as words or phrases as part of controlled vocabularies; ‘concepts’ are defined as ideas directly relating to content; and ‘related concepts’ are defined as other ideas relating to concepts already selected. Searching by ‘key word’ was most popular with 181 votes (44%) followed by ‘subject terms’ (133, 32%), then ‘concepts’ (66, 16%) with the least preferred option being ‘related concepts’ (33, 8%). Table 3-5.

Table 3-5 Preferred search options

	Responses	
	Number	Percent
Keyword	181	44%
Subject terms	133	32%
Concepts	66	16%
Related concepts	33	8%
Total	413	100.0%

Survey respondents were also asked to indicate which aspects of a research study should be easily searchable. Results show that the ‘research study question’ was the most popular answer with the respondents followed by ‘variables’ and then ‘research publications’ (Table 3-6).

Table 3-6 Searchable aspects of a research study as indicated by role

Role	Searchable aspects of a research study									
	Research study question	Variables	Research publications	Research study protocol	Data collection instrument designs	Code lists	Research data management plan	Funding details	Concepts	Consent form and associated information pack
Data user	124	116	113	111	101	87	48	43	34	39
Data provider	91	89	80	80	75	69	36	30	28	36
Archivist / Librarian	17	17	17	12	15	17	3	5	15	4
Policy maker	11	10	10	11	8	10	4	5	3	6
Other	12	8	9	11	10	6	6	7	8	5
Funding agency	5	5	5	5	5	4	2	4	2	3
Observer	1	1	1	1	1	0	1	2	1	1
Total	162	152	147	143	129	116	61	55	55	53

Finally, the respondents were asked to indicate which of the specified list of terminologies, classification systems, thesauri and metathesauri they were familiar with. The most familiar standard was the International Classification of Disease and the least familiar was the European Language Social Science Thesaurus, Table 3-7.

Of the 9 who specified they used other tools, two examples of responses include, but are not limited to, the WHO Anatomical Therapeutic Chemical Classification System with Defined Daily Doses (ATC/DDD) codes and WHO Drug Dictionary. The respondents were also asked to provide details of any tools they used to assist the management of controlled vocabularies, responses include the WHO Global Health Observatory.

Table 3-7 Classifications and thesauri

	Responses	
	Number	Percent
International Classification of Disease (ICD)	148	32%
Medical Subject Headings (MeSH)	119	25%
Diagnostic and Statistical Manual of Mental Disorders (DSM 5)	78	17%
SNOMED CT	36	8%
Read Codes	34	7%
OPCS-4	17	4%
Other	9	2%
Humanities and Social Sciences Electronic Thesaurus (HASSET)	11	2%
European Language Social Science Thesaurus (ELSST)	4	1%
Unified Medical Language Service (UMLS)	5	1%
Logical Observation Identifiers Names and Codes (LOINC)	6	1%
Total	467	100.0%

3.4.2.4 Data documentation and metadata standards

The survey also aimed to identify current challenges in creating and/or using data documentation (a total of 50 people answered this question). The challenges fall into the following categories:

1. Standardisation
2. Resource availability
    a. General comments
    b. Time and costs
    c. Technology
    d. People/staffing

A total of 16 people commented on standardisation with 9 focusing on data management and 7 commenting on research data. Of those which commented on standardisation in data management, a common concern was the current lack of data management standards with one person stating there was an,

> "…inappropriate data management tools (and) Poor standardisation of data dictionary".

Two people commented on marking up variables and

> "ensuring sufficient documentation on derived variables"

with a third noting

> "providing sufficient detail and access while retaining data confidentiality"

is important. A common concern held by those who commented on standardising the data itself, was how the heterogeneity of the data impacts documentation. One respondent commented,

> "creating data is easy, documenting data is the rather more complex issue"

whilst another commented on

> "The complexity of data...collected using CAPI is a challenge"

and representing,

> "an instrument that may have been experienced differently by different respondents to researchers in a straightforward way is difficult."

A total of 29 people commented on issues relating to the availability of resources to cover the time required, financial cost and availability of trained staff to document data. Generally, it was felt by respondents that more resources are needed to support data documentation. One respondent stated that,

> "skill, time, resources and the less tangible aspect of lack of recognition/appreciation for doing this work…one does not normally get grants, get REF brownie points or publication kudos out of producing this documentation."

In terms of time and costs, 9 people commented specifically on the lack of time and funding for data documentation with one respondent saying, "I would need funded time for data management staff" to complete this task. With regards to technology, a total of 4 people commented specifically on use of technology as a resource with a "lack of good quality tools" and standards being two such concerns. Lastly many of those who responded commented on the implications generating data documentation has on people and staffing. Further to time and costs, skills development; or rather, a lack of sufficiently skilled staff is a particular challenge. This subsequently raises issues over current skills development programmes and whether more can be done in the way of training.

The most commonly indicated metadata standard were the Data Documentation Initiative 2/3 standards, particularly with data providers employed by universities, Supplementary Table 2.

In terms of tool availability, this ranged from in-house options such as those developed by WWARN to commercial tools such as Colectica. Another commonly used tool was the Nesstar publisher. Other responses included

Microsoft Excel and Word with one respondent writing they used “NVivo, Zotero” to assist data documentation.

3.4.2.5 Promoting data discoverability and use of data repositories

The most popular repository was ClinicalTrials.gov (23%). Of the repositories respondents were intending to use, again ClinicalTrials.gov was the most popular (7%) and the least popular was social science (0%), Table 3-8

If the option ‘Other’ was selected, the respondents were then given the option of providing details of any other repositories not listed, exemplar answers included, the “INDEPTH Data Repository” and the “Spanish National Institute of Statistics”.

The respondents were also asked to provide their opinion on areas of importance to data discoverability (Table 3-9). The numbers in bold signify where the median is.

Table 3-8 Current and intended use of repositories

	Use of repositories					
	Already used			**Intended to use**		
Repositories	**Number of respondents**	**% of survey respondents**	**CI**	**Number of respondents**	**% of survey respondents**	**CI**
ClinicalTrials.gov	58	23	22-36	18	7	4-17
Social Science	47	19	18-30	1	0	0-67
Other	20	8	6-15	6	2	3-15
Genetic association & genome variation	17	7	5-13	5	2	2-14
Environmental & geoscience	16	6	5-13	9	4	1-12
Organism or disease specific resources	13	5	3-11	7	3	4-17
Functional genomics	8	3	2-8	4	2	1-12
DNA protein sequences	6	2	1-6	2	1	0-9
Figshare	5	2	0-6	9	4	1-12
Molecular interactions	4	2	1-5	2	1	0-9
Molecular structure	3	1	0-4	2	1	0-9
Taxonomy & species diversity	3	1	0-4	5	2	2-14
Dryad	2	1	0-4	6	2	3-15
Proteomics	1	0	0-3	5	2	2-14
Total	203	80		81	32	

Table 3-9 Aspects important to discoverable data

Aspects of importance to discoverable data	Opinion					
	Essential	Extremely	Fairly	Slightly	Not at all	*(blank)*
Be on the web						
Number of survey respondents	83	69	31	7	9	54
% of survey respondents	33%	27%	12%	3%	4%	21%
Be provided in a machine-readable format						
Number of survey respondents	94	74	19	3	6	57
% of survey respondents	37%	29%	8%	1%	2%	23%
Be provided in a non-proprietary form						
Number of survey respondents	44	61	64	5	14	65
% of survey respondents	17%	24%	25%	2%	6%	26%
Conform with recognised data management standards						
Number of survey respondents	55	82	50	1	5	60
% of survey respondents	22%	32%	20%	0%	2%	24%
Be linked to an underlying conceptual framework or ontology						
Number of survey respondents	20	45	78	14	34	62
% of survey respondents	8%	18%	31%	6%	13%	25%

3.4.2.6 Data publication and citation

The survey showed that most common way of first hearing about data publications was through a colleague (35%). This was then followed by journal (29%), conference/workshop (23%), search engine (6%) and the other option (7%), to which respondents offered “Involved in projects in this area”, “Blog posts” and, “I edit a journal where we promote publication of data resource papers” as three such alternative methods.

The respondents were also asked to indicate which of the perceived benefits of citation (adapted from a report by Ball and Duke (2012) from the Digital Curation Centre) they consider to be of most importance. Respondents were able to make multiple selections. It was identified that the ability to ease the process for readers to locate the data and attributing credit to those who contributed the data was the two most important benefits of data citation, Table 3-10.

Table 3-10 Benefits considered important to data citation

	Responses	
	Number	Percent
Easier for readers to locate data	137	22%
Proper credit given to data contributors	113	18%
Links between datasets and associated methodology publication provide context for reader	114	18%
Links between datasets and publications describing their use can demonstrate impact	77	12%
Infrastructure can support long-term reference and reuse	69	11%
Promotes professional recognition and rewards	64	10%
Less danger of data plagiarism	40	6%
Other	4	1%
Total	618	100.0%

Other benefits of data citation (provided under the ‘Other’ option) included “avoid duplication of effort e.g. for validation studies” and “Provides accurate and sufficient wording to use in reference lists”. A further two benefits were provided, the first discussing the potential for these to be "…assessed for its own value, provenance and potential…" and the second as a mechanism to help,

> "…secure funding from agencies such as the National Institutes of Health and the National science Foundation for new and continuing data integration and dissemination projects. We must prove that the datasets are being used for science and policy and can only do so if users appropriately cite our datasets."

Continuing this focus on data citations, respondents were asked to indicate how granular data citations should be. Results showed that the most popular level of citations were at the data collection level followed by a single dataset (or sweep). For those that selected the 'other' option, an exemplar answer included, but is not limited to "subset of data used for analysis" (Table 3-11)

Table 3-11 Granularity of data citations

Levels of data citation	**Percent**
Dataset collections	36
Single datasets (or sweep)	33
Files within datasets	16
Individual items of data	13
Other	2

To help identify best practice in managing longitudinal and regularly changing datasets, the respondents were asked to indicate how these data should ideally be handled. The most common approach was to 'Publish revisions at regular intervals' followed by 'All published versions should be published in instalments', Table 3-12.

Two suggestions were made for alternative approaches; the first was to include lists of, "…changes made from release to release…Also provide any crosswalks to enable linking between different releases/time periods." The second was to make use of 'timestamp' to assist identification of changes.

Table 3-12 Managing longitudinal and regularly changing datasets

Approach	Percent
New identifier assigned at each update	18.3
Publish revisions at regular intervals	28.9
Time series data should be published as complete 'snapshots'	16.3
Time series data should be published in instalments	10.3
All published versions of the datasets must be stored	25.1
Other	1.1

The survey identified a number of challenges currently associated with the adoption of data publications. These can be categorised into the following:

a) limited significance: A number of respondents questioned the value of data publications with several stating that there is little incentive to publish, for example, the publications are not acknowledged in the REF. One respondent felt these articles were unnecessary:

> “In many fields, the information in a data publication is already available in documentation accompanying the data”

b) Inadequate resource availability: Time and costs were the most commonly discussed with their efficient use being a concern. One respondent wrote these publications are “resource intensive…” and another described them as, “…likely to be lower priority than publishing actual research on the data”. Another wrote that there is,

> “excessive linking of university fuinding and employment and tenure to outdated private journal models based on prestige and less accessible to non-established researchers and in developing countries”.

c) Need for changes in research culture: six respondents each commented on research culture and current approaches to scientific research. One respondent commented,

> "Culture change –specifically, that data be considered and acknowledged (by funding bodies and employing institutions) as a valuable scholarly output alongside publications".

It was also noted that researchers are "…over-stretched…" and there is "…a shortage of expertise in data management staff."

These findings highlight the limited perceived significance of data publications and the undervaluing of efforts to produce these particularly within the context of academic reviews where HEFCE block funding could be awarded. Additionally, the perceived lesser contribution these make to science in comparison to original research articles appears to discourage researchers from prioritising their creation. These results also suggest that funding agencies and institutions could do more in the way of promoting data publications and in particular provide greater support for researchers to produce these. Consequently, the use of data publications as a mechanism to enhance the discoverability of research data may be negatively affected.

3.4.2.7 Areas of importance to data discovery

Areas of importance to data discoverability are categorised as follows:

Identification of commonalties and links between studies: The ability to identify common research questions and hypotheses will help reduce redundant research efforts and help streamline research processes. One respondent wrote that it was, "important to encourage data analysts to move out of the comfort zone of using a favourite study and to use several complimentary sources" However, the ability to make links between studies is determined in part by the extent to which they are published which is a "…major challenge (in) unpublished data, particularly for trials…" The process of obtaining data is another issue which needs to be addressed through, "…simple…" processes which are not subject "…to constant change".

Metadata markup and producing other associated documentation: Producing metadata and other study artefacts will help secondary researchers identify "…any local influences on variables/interpretation" as written by one respondent and "…ensure transparency and reproducibility" as written by another.

The request was also made for researchers to be able to search for particular instruments and, "…to discover subject characteristics, such as gender, age, socio-economic status". Provision of a "detailed analysis plan" will also support data discoverability. Regarding use of schemas, the argument was made for "a clear and transparent schema…" for variable name annotations as this would be "most helpful" if these accompanied the questionnaires.

Use of technology: The use of "…new, often complex…" databases were argued by three people as an important aspect of data discoverability as,

> "public health research depends on breaking data silos, which is not possible with conventional querying procedures on conventional databases."

In terms of data warehousing, one person suggested, "development of data warehousing standards and practices for medical data". Additionally, the capacity for data linkage and/or harmonisation was raised by two respondents. One discussed the capability of linking data from,

> "…sources such as administrative data bases" whilst the other questions whether in terms of "harmonization – has the data set already been transformed into a simplified standard format?"

Consent and other ethical issues: Three people raised the subject consent respecting the rights of the participants'. More specifically, "…meeting privacy/confidentiality requirements" using,

> "standards for releasing protected health and public health data…" and gaining "…public trust need to be of the highest priority".

These findings suggest that greater focus is needed on enhancing the infrastructure supporting epidemiological and public health studies and in particular, the way in which research data are managed. The results suggest that standards which encourage the uniform development of data warehouses and mechanisms to enable data linkage and/or harmonisation are required to facilitate this enhanced management, particularly so that the boundaries of consent are respected. Greater focus on enhancing the infrastructure could also potentially encourage researchers to produce more detailed metadata to assist the management process. This is important to data discovery as the metadata can provide the details needed to understand the data, the circumstances under which data were collected/generated and the potential for data integration creating even richer datasets.

3.4.2.8 Areas of improvement and immediate priority

A total of 73 people answered the question focusing on areas of improvement and immediate priority. The following areas were suggested:

a) registration of studies: a total of 11 people made the argument for some kind of public health portal/register whereby details of studies can be uploaded with two citing ClinicalTrials.gov as an initiative whose approach to recording studies should be taken into consideration. Having a single point of discovery encourages development of standardised study documentation and could enable "cross collection searching…" Furthermore, this kind of approach to improving discoverability has the potential to enable researchers to define data access and sharing policies, caveats and conditions for use. There appeared to be no expectation by the respondents that the data itself be made available but certainly a record confirming the data's existence was firmly encouraged.

b) Standards: A number of people raised the issue of increased use of standards. One respondent discussed the,

> "need international standards (not just European vs. rest of world). Basic information about the type of information that is collected and surveys administered for each country needs to be made available, not just in native language."

Another respondent discussed the need for "greater harmonisation of indicators…and more standardised formats for storage and access". Three respondents suggested international agreement of standards and approaches. Specifically, "a global framework of best practice", "international criteria developed for data sharing…" and a "…universal strategy…" were requested by the respondents.

Another individual discussed use of the,

> "…ISO 13606 family of standards, the openEHR or the Multilevel Healthcare Information Modelling (MLHIM) specifications…"

Having a standardised working environment promotes,

> "…semantic interoperability between distributed, independently developed databases, which solves all the technical aspects of data discoverability…"

This particular individual then suggested use of Semantic Web technologies to assist implementation of these kinds of standards.

Five people raised the issue of metadata standards and their increased application in the field. For example, one respondent discussed use of

> "…internationally recognised data documentation standards e.g. DDi lifecycle" and another suggested "Public health data should be curated to appropriate metadata standards…"

Another respondent said there is a,

> "…lack of true assessment and straight reporting about how relevant studies are to the original populations and how generalizable findings might be".

c) Publications and citations: Improved data publications and citations were raised by several people. One respondent called for funders to link publications to the datasets they financially support whilst another called for "improved data citation practices…" The use of identifiers and in particular digital object identifiers was also raised in the context of publications.

d) Consent and other ethical issues: ensuring the wishes of participants are respected and mechanisms are in place to protect participants' identifies were suggested as areas of improvement. One respondent also noted a "…lack of confidence by ethics committees and data custodians" and that this negatively impacted data access and linkage.

These findings suggest that more is needed in the way of standardisation to address the current insufficiencies in the way in which research data are made discoverable. More specifically, increased use of standards have the potential to improve the quality of data and provide the basis from which standardised research data descriptions may be created.

3.4.3 Mechanisms to enhance discoverability

1. Data publications

This is the provision of collections of data publications which provide detailed descriptions of datasets. The datasets themselves do not need to be publicly available but any unrestricted metadata or other associated artefacts should be made publically available. Where possible, these publications should provide links to where the data themselves may be accessed and include details of any caveats and data use/access policies so as to inform potential secondary users and help characterise the datasets. It was through the systematic literature review that I identified the potential to use these publications as a mechanism to enhance data discoverability.

Currently, several journals already publish data papers; for example, the International Journal of Epidemiology publishes 'Cohort Profiles' each describing epidemiological studies. This particular journal also publishes 'Data Resource Profiles'. These two types of publications differ in that 'Cohort Profiles' describe a particular study of the cohort of participants; whist, 'Data Resource Profiles' describe a platform/source of data. These are another type of data publication but are specifically for describing epidemiological data for purposes of testing hypotheses and performing analyses. Both types of data publication are concise (2500 to 3000 words) and are indexed by biomedical databases such as PubMed. Other journals include Scientific Data as published by the Nature Publishing Group. This multidisciplinary journal publishes 'Data Descriptors' which describe pre-existing datasets for purposes of promoting data discovery and reuse. Ubiquity Press also publish a series of open data journals such as, Journal of

Open Health Data (2016k), publishes papers describing datasets with reuse potential and the Open Journal of Bioresources (2016m) which publishes papers describing bioresources (collections of biological data and samples) also with reuse potential. The advantage of publishing data or bioresource publications in journals such as these, the papers can be cited correctly and there is the potential to track the level of reuse.

2. Sematic Web technologies

This model utilises SWTs maintained by the World Wide Web Consortium (W3C 2014b), to link data and metadata together which can then be published on the Web. SWTs can be used to enhance data discoverability as standards such as the RDF can be utilised to build searchable networks of epidemiological and public health information online. The scalability of the RDF is such that these networks can build without the potential to negatively impact existing links. The use of OWL ontologies can help to structure this epidemiological and public health information in a way that is controlled and yet has the potential to grow larger as more information is added. It was through my research in this thesis (chapter 2) that I identified the potential to use SWTs to further enhance data discoverability.

Organisations such as the Semantic Web Health Care and Life Sciences Interest Group are already looking at ways to embed SWTs into epidemiological and public health research to harness their potential benefits to the research process.

3. Public health portal

This involves the use of a public facing catalogue containing records of pre-existing public health and epidemiology studies. The centralised portal should be freely available online and where possible, provide links to where more information can be found relating to the individual studies. I identified the potential to use a public health portal to enhance data discoverability through findings from my qualitative analyses of the survey results. I also identified the need for a registration process for observational studies which could potentially be enacted through the use of a public facing portal.

There are portals such as these already publically available such as ClinicalTrials.gov. Another such example is the CALIBER portal (Denaxas, George et al. 2012). Researchers using CALIBER data are required to provide details of their study, with the ClinicalTrials.gov identifier, which is then made available through a public facing catalogue. Variable-level detail is also provided on the portal. Other examples include the CESSDA(2016e) portal and the UK Data Service(2014h).

3.4.4 Evaluation of mechanisms to enhance discoverability

I performed a series of feasibility analyses and engaged with stakeholders at the Public Health Research Data Forum to evaluate the mechanisms.

3.4.4.1 Data publications

Technical: Data publications are an established form of literature with journals such as Scientific Data (Nature Publishing Group) providing a mechanism through which papers describing scientific datasets may be published. To uniquely identify data publications globally, specialist technical support would be needed to mint DOIs; the DOI would resolve to a landing page for the publication. This too requires additional resources as for an organisation to mint and maintain DOIs, they must have the finance infrastructure available to maintain the DOIs and landing pages in the long term.

Organisational: The submission process of these manuscripts utilises journal submission webpages and are likely to be subject to peer review. Stakeholders are likely to be familiar with the submission and publishing process as it very much mirrors that of research articles. Familiarity with this kind of publication, and their potential benefits is increasing, however formal academic recognition of their significance is limited. Review processes such as the REF do not acknowledge their value and consequently stakeholders are not incentivised to produce such publications. However, stakeholders are likely to need increased support when producing such publications. Guidance for authors is generally provided on journal websites but writing data publications concisely with the appropriate level of detail can be time-consuming and potentially increase the workloads of stakeholders.

Economic: data publications can have an associated processing charge or publication fee. Journals such as the International Journal of Epidemiology (Oxford University Press) offer different price models depending on the location of the author and the type of access ranging from £0 to £2000[4]. Data publications are potentially costly but varying cost models can help researchers in publishing their manuscripts. Another factor to take into consideration are the changes taking place around the way in which publications are made openly accessible. As from April 2016, HEFCE mandate that the author accepted manuscript must be made openly accessible within 90 days of acceptance. There are two ways to do this: a) green route whereby the publication is embargoed; or b) gold whereby the publication is made immediately openly accessible but potentially at a cost. Furthermore, it is unlikely that additional software or tools are likely to be needed/purchased when producing data publications – commonly used word processing packages should suffice.

3.4.4.2 Semantic Web technologies

Technical: Harnessing the benefits of linked data and SWTs is advantageous as methods such as the RDF use uniform resource identifiers (URIs) to provide a rigorous method to differentiate between resources thus supporting data integration processes (Pathak, Kiefer et al. 2012b). However, any changes made to the URIs must be managed effectively otherwise the links could incorrectly resolve. Producing high quality and responsibly managed URIs remains a challenging yet intrinsic part of integration processes. Another potential issue with the implementation of RDF technology is meeting the need to develop and execute queries using disparate sources of data (Pathak, Kiefer et al. 2012b). The stability of SPARQL endpoints in addition to the current need for user-friendly, reliable

[4] correct November 2015

tools to write SPARQL queries could limit the extent to which these technologies are implemented and their potential benefits realised (Pathak, Kiefer et al. 2012b).

A number of RDF conversion tools are available, a list of which can be found on the W3C's website (W3C 2014a). One example is Bio2RDF as supported by the W3C's HCLS IG. Tools such as these can help researchers and clinicians to mark-up up clinical data potentially increasing their interoperability.

Another mechanism which could potentially be utilised as part of this model is ontologies. An example of a biomedical ontology is the TMO developed by the HCLS IG's Translational Medicine task force (Luciano, Andersson et al. 2011). The TMO as part of the Translational Medicine Knowledge Base (TMKB) is an initiative which represents drug data using the RDF and maps to over 60 other ontologies (Luciano, Andersson et al. 2011; Pathak, Kiefer et al. 2012b). Research into mental health disease has also benefited from application of SWTs. The HCLS IG's Semantic Web Applications in Neuromedicine (SWAN) ontology was developed to help improve knowledge management of Alzheimer Disease and connect resources together (Ciccarese, Wu et al. 2008).

Ontologies may be collected together and presented in public facing catalogues. The BioPortal (2014a) is an open repository which collates the ontologies written using OWL, RDF, OBO and Protégé and their associated metadata (Noy, Shah et al. 2009). A key feature of this tool is the provision of mappings and mapping annotations (metadata) between the ontologies to facilitate integration endeavours; it is possible to link records together using common ontological annotations (Noy, Shah et al. 2009; Whetzel 2013). The National Cancer Institute(National Cancer Institute 2014) use the BioPortal as a repository for their clinical ontologies (Noy, Shah et al. 2009). In having biomedical ontologies available in public facing catalogues, it is possible to see how other research groups have structured their clinical data and there is the potential to reuse the ontologies. In being able to use pre-existing ontologies, scope for collaboration increases and the possibility of duplicate efforts potentially reduced.

Organisational: familiarity with linked data on the World Wide Web and use of methods such as RDF and OWL is currently limited potentially causing a significant demand for training. Also, integrating use of such tools into daily work routines would require effective change management so as to minimise any potential disruption experienced by stakeholders.

Furthermore, this model is potentially biased towards those with internet access based in higher income countries. This is a weakness of the model as those from lower and middle income countries, where internet access can be variable, may be at a disadvantage. Consequently, they may not fully benefit from the potential research opportunities associated with use of SWTs in epidemiological and public health research.

Economic: A number of open-source software packages are currently available to the scientific community. For example, Bio2RDF is an open-source tool designed to integrate data from over fifty sources such as NCBI's Entrez Gene and OMIM with release 2 having a total of one billion triples using SWTs (Belleau, Nolin et al. 2008). Protégé, developed by researchers at Stanford University is a free open-source ontology development platform which can be utilised to create OWL ontologies(2015d). Nevertheless, use of any commercial RDF convertors or ontologies development platform could cause stakeholders to incur purchase and license fees.

Additionally, given the need to maintain equity, financial costs could be incurred when implementing the necessary infrastructure in locations where this is currently insufficient. There is also the need for increased training and support, both of which may have an associated cost. Therefore, factors such as these could negatively impact how widespread adoption of this model is and its continued use.

3.4.4.3 Public health portal

Technical: To build and implement a public health portal, there are a number of stages each with their associated technical factors. The first step is to gather stakeholder requirements to inform design and development of the portal. This is potentially complex task given the diverse nature of epidemiological and public health research the tool could potentially encompass.

The second step is to build an underlying metadata model consisting of several layers to structure the metadata. The first layer would consist of general information potentially utilising the schema for Simple Dublin Core. This is a multidisciplinary metadata standard which enables generic metadata to be recorded. The second layer would depend on which sub-domain of epidemiological and public health research a research study sits. This would firstly utilise some kind of ontology to structure this knowledge effectively and provide a mechanism through which changes to the different categories of research may be made. Engaging with stakeholders will help to define the ontology; although this is potentially problematic given the possibility of a research study to sit in multiple areas. Once the ontology has been developed, a standard such as DDI may be incorporated into the metadata model to record lower level detail. The advantage of using metadata standards such as these is that DDI has the Simple DC schema built into it (Figure 2-3) so the use of both these simultaneously is potentially straightforward. Another advantage is that linking to other portals such as the MRC Gateway and UK Data Archive is potentially enabled as their metadata is DDI compliant.

The third step is to build a test portal to establish whether all user requirements have been addressed. It is at this stage that strengths and weaknesses of the model can be identified and corrective action taken if necessary. The fourth step is to build a pilot portal where stakeholders from across the research domains can be engaged to determine whether the portal is fit for purpose.

Organisational: use of online portals such as ClinicalTrials.gov and CESSDA are an already established practice in epidemiological and public health research. Therefore, stakeholders are likely to be familiar with the use of online portals to support research efforts.

Nevertheless, the challenge will be integrating the use of a novel public health portal for observational studies into stakeholders’ work routines.

A potential mechanism to ensure pre-existing and newly established studies register with the portal would be for it to become a funder requirement.

Economic: Initially, resources are needed to perform requirements gathering and building of a test portal. Once strengths and weaknesses have been identified, and changes made, further funding is needed to support the development of a pilot portal containing test cases (which can be evaluated) before a finalised portal may be developed and become available for use. In the longer term, funding is likely to be required to maintain the portal and records, and to allow enhancements to be made where necessary.

3.5 Discussion

3.5.1 Systematic literature review

From this review, the use of methods to enhance data discoverability is limited. For example, three studies (ID's 3, 6, and 13) do not provide information around the accessing of data. The review also found inconsistencies in the way in which research data were documented as described below. Therefore, further investigation into the issue of data discoverability is needed with a view to identifying ways in which the discoverability of research data may be improved.

This review identifies a range of approaches to providing data documentation; for example, three studies (study IDs: 18, 33 and 48) provide data dictionaries while another provides questionnaires (study ID: 13) and a third providing a manual (study ID: 22). The review also identified application of the metadata standard, DDI in three of the studies (study IDs: 18, 23 and 48) all of which make use of some kind of catalogue to provide data documentation. Therefore, the use of metadata standards and the process of searching through catalogues to find metadata in addition to other approaches to identify research data e.g. data papers (such as those published by the International Journal of Epidemiology) were investigated further as part of the survey.

The review was also successful in identifying that all the studies provided links/descriptions of related publications and 5 studies (study IDs: 14, 18, 20, 22 and 49) provided tools online for data visualisation and analysis; both of which can help potential secondary users characterise the data before requesting access. A total of 11 out of the 13 studies (study IDs: 6, 14, 18, 20, 22, 23, 26, 33, 34, 48, 49) provided descriptions pertaining to data access. Therefore, issues relating to access and in particular, use of metadata repositories will be investigated further to help further characterise current practice.

A potential weakness of the review is the way in which it is structured. If the review were to be repeated, the studies would be categorised according to the amount of funding they receive; for example >£10000, £10000-£20000 etc. This is because the differences in approaches to discoverability could be attributed to the level of resources a study has – stakeholders may want to use a range of mechanisms to facilitate discoverability but only have resources for one. By grouping the studies according to funding, this would potentially give a clearer overview of what can realistically be achieved using a finite amount of resources. This issue of available resources is an area in need of further exploration and could potentially have an impact on the models suggested – a study cannot be penalised for not having enough resources.

3.5.2 Online stakeholder survey

Data types: Most of the research studies identified through the systematic literature review were observational which could probably account for why the three most popular forms of data were ‘survey’ followed by ‘healthcare records’ then ‘disease registries’.

The research data lifecycle: The results of survey illustrate less activity towards the end of one iteration of the research data lifecycle. This can probably be attributed to the nature of the jobs held by those who completed the survey. The most common role in public health research data was data user; this is reflected in the results relating to the area in which most of the respondents were actively involved in - analysis. The next most popular stage in the research data lifecycle is access, use and reuse; again these results correspond to the role in public health; the second most commonly cited role was data provider. The stage in which the fewest number of people indicated they were actively involved was data destruction. With current drives to improve data reuse and enhance use of resources, the destruction of data does not seem in keeping with current data reuse and repurposing initiatives. However, again, job role could explain these results. For example, in the event of a participant requesting their data be destroyed, then this must occur to the best of the ability of the person(s) responsible. (Acknowledgement is given to the inability to destroy individual-specific data from anonymised aggregate datasets, especially if these data have been shared with third parties). Therefore, the respondents who indicated they were involved in data destruction could form a part of this particular group of stakeholders within the wider scientific community. Further investigation is needed to confirm this, and the ways in which data are destroyed, to establish how these individuals may be better supported. These particular stakeholders are fundamental to the process of respecting participants' rights and the limits of their consent – consent and other ethical issues is a recurrent theme in the results of this study.

Search options: When asked to indicate preferred search options, the most popular methods were 'keyword' followed by 'subject terms'. Given key words can be sourced from classification systems such as Medical Subject Headings (MeSH) or the International Classification of Disease (ICD); it is possible that since the survey was aimed at the research community in particular, whereby use of classification systems such as these is commonplace, these proved most popular in the survey results. The search option results correspond with the results of the controlled vocabulary and thesauri question asking participants to indicate which controlled vocabularies and thesauri they are familiar with – ICD and MeSH were the two controlled vocabularies most familiar to the survey respondents. In terms of future work, the next step would be to investigate at a lower level the extent to which 'key words' were selected from controlled vocabularies and the possible implications of this practice.

Controlled vocabularies: Datasets such as Hospital Episode Statistics (HES), which can be used as part of public health and epidemiology research, provide clinical information using ICD and OPCS.

Given that the survey was aimed at stakeholders from the public health and epidemiology research communities, this result could be a reflection of the daily work routines of those who completed the survey. If the survey had been aimed at clinicians, it is possible that SNOMED CT could have scored more highly.

Thesauri: The controlled vocabulary respondents were least familiar with was the European Language Social Thesaurus (ELSST). This thesaurus' English translation however, the Humanities and Social Sciences Electronic Thesaurus (HASSET), scored slightly higher with 11. This is likely to be because ELSST (and HASSET) are used primarily in the social sciences; whilst the survey was aimed at stakeholders within the public health and epidemiology domains. It is possible that had the survey been aimed at social scientists too, both these thesauri may have scored more highly. Further analysis and in particular a widening of the scope to include those in the social sciences and clinicians is needed to better understanding of the use of controlled vocabularies and thesauri in research to inform application in clinical environments.

Data documentation and metadata standards: Following collection of results, the low level of response to the question asking which of the specified metadata standards respondents had knowingly used, could be attributed to the nature of the job roles of those who completed the study. The majority of respondents indicated their role in public health research data were as users. Therefore, since metadata markup may not necessarily form an everyday part of data users daily work routines, this could explain the low level of response. Had the results included more individuals from the archival and librarian communities, the number of respondents could potentially increase. All questions were optional so this too may have been a factor contributing to the low level of response.

Nevertheless, with study artefacts such as protocols and data dictionaries often being provided in formats such as PDF (based on results from the review conducted earlier in the study) this could also attribute for the low response rate. The artefact could in fact be standardised but the researcher might not necessarily be aware of this unless they request a machine-readable format of the documentation such as XML. Regarding tool availability, even though several examples were provided by the respondents, some in-house and some commercial, more work is needed to establish how mature these tools are and implications of their use; for example, are there any potential costs or specific training needed to enable use.

With current motivation to increase use and standardisation of metadata these results show that more emphasis is needed on the importance of metadata as supported through increased training and support for researchers using the metadata (and data) and the data managers and other professionals who are responsible for the metadata and data. Further investigation will enable identification of the points in the metadata pathway from generation through to sharing which require further support and/or improvement.

Data repositories: Use of specialised repositories is essential given the intricacies of biomedical data and the need to ensure they are correctly managed. The most commonly used repository, both already used and intended for use was ClinicalTrials.gov. Registration with this repository is compulsory for clinical trials so this could account for the high proportion of votes. The second most popular repository already used was Social Science. This could be because organisations such as the UK Data Archive are responsible for the management of a number of datasets which have the potential to be used as part of public health and epidemiology research. An emerging common theme was the use of metadata repositories, examples of which include the “MRC Research Data Gateway” and the “Australian Institute of Health and Welfare Meteor metadata repository.” Additional investigation could help determine the extent to which metadata repositories are used by in public health and epidemiology research and how those using and those maintaining them can be better supported. This too will help inform development of best practice guidelines and gold standards.

Data publications: The most commonly indicated method of knowing about data publications is through a colleague followed by a journal. Based on the responses from the open-ended question focusing on challenges associated with the widespread adoption of data publications, three key themes emerged: significance, resource availability and research culture. Further work focusing on the promotion of data publications, particularly as a method of charactering datasets as a mechanism to enhance discoverability and to alter perceptions relating to their significance is needed.

Data citation: Regarding the benefits of citation, being 'Easier for readers to locate data' was most commonly indicated by the respondents; the least commonly cited were 'Less danger of data plagiarism' and then 'Other'. Based on these results, and suggestions provided by the respondents, it is clear that many researchers view data citation as being particularly important to the attribution of credit and supporting increased visibility of datasets.

3.5.3 Evaluation: Mechanisms to enhance discoverability

Data publications may be used by researchers to publicise their data to improve discoverability; whilst, retaining a sense of ownership of the data and respecting the participants' rights (de Carvalho, Batilana et al. 2010). There are established models and academic journals for publishing this kind of manuscript and they are already being indexed in biomedical literature databases such as PubMed.

Furthermore, there is opportunity to assign metrics to the publications for researchers to monitor the number of times their paper has been read, downloaded and/or cited. However, caution is needed so as to not turn data publications into a way of monitoring the publication success of researchers or as an alternative to data sharing. The purpose of data publications is to provide unambiguous descriptions of research data and the process through which they are/were managed.

There are also potential problems associated with the enhanced promotion and adoption of this model. For example, stakeholders in public health and epidemiological research must be sufficiently trained and experienced to produce such publications. There also needs to be adequate resources available such as time and finance to support the writing and publishing processes. Additionally, stakeholders are discouraged to produce data publications due to their perceived limited significance given the lack of formal academic recognition.

Data publications can work in both the short and longer term. They offer stakeholders a mechanism to describe their data without infringing on any confidentiality or privacy laws/ governance frameworks if written well. Of the three models, this is the most organisationally demanding as changes are needed to current research culture. Much is still needed in the way of increasing their prominence in scientific publishing and improved academic recognition is sorely needed.

The second model was use of linked data on the World Wide Web. The potential benefits of having a linked data approach through application of SWTs have already encouraged discussion within the clinical research community (Sinaci and Laleci Erturkmen 2013). Having metadata could potentially improve the efficiency of searches online and help to improve identification of resources (2009).

Additionally, RDF networks could include links to social media sites and other such potential providers of metadata. The review identified examples of where the benefits associated with social media have been harnessed to enhance data discovery. Having links to metadata on websites such as these could enable creation of metadata tailored to a particular audience. For example, interviews with principal investigators may be uploaded to these sites describing the research study aimed specifically at the general public. Being able to facilitate public engagement is important to a research study and this model could lend itself well to enabling this.

This approach does however pose several challenges. For example, inadequate provision of machine-readable formats of the data and metadata, such as RDF, limit the model’s application. Transforming the information and continued maintenance requires additional resources as does ensuring unique identifiers resolve to the correct resource (Belleau, Nolin et al. 2008; Katayama, Wilkinson et al. 2013). Furthermore, by applying SWTs to create RDF networks to provide linked data on the World Wide Web; there is the potential for URL links to resolve incorrectly, if at all. The ability for networks to be built up without some kind of control is an inherent weakness of the linked data model. Nonetheless, use of linked data in life sciences research is increasing, and the potential benefits of applying formalised data management techniques such as ontologies from computer sciences in medicine are already being realised. In 2.5.1, detailed examples of where SWTs have been applied in life sciences research are provided.

Widespread use of linked data on the World Wide Web and SWTs is very much a longer term goal of enhanced data discoverability. Of the three models identified, this model is the most technically demanding. Implementation of this model is inherently challenging given the high dependence on the technical expertise of stakeholders and ready access to the necessary infrastructure. Much has already been achieved using SWTs (as described in chapter two) yet more is needed in the way of championing the potential research benefits associated with this model to stakeholders.

The third model was the public health portal. The potential for a single portal to contain records of public health and epidemiology studies would make a valuable research tool. This, joined with recent computational advances enabling more affordable analysis of big data(Chute, Ullman-Cullere et al. 2013) and an increased capacity for data linkage, as supported by robust mechanisms to help protect participant privacy and confidentiality,(Lyons, Ford et al.) serves to enhance the research working environment.

Currently there is no known mandatory registration process for observational studies which could be a contributing factor in the limited discoverability of certain research data. By having a standardised registration process for these studies in addition to having the metadata publically available, scope for data discovery and subsequent reuse is potentially increased. Furthermore, having this kind of registration process also facilitates application of standards to the metadata. There is the potential to use a metadata standard such as DDI to structure the underlying XML and opportunity to use encoding standards such as ICD to populate the metadata fields.

Nevertheless, development of the portal raises concerns over its continued maintenance, study registration (additional analyses possibly required to determine feasibility), and the sustainability of this model i.e. could the portal grow too big? Issues are also raised relating to stakeholders having the time to enter details of their research data and assigning the task of monitoring the records to a member(s) of the research team. There is also the challenge of addressing any language barriers which may arise. Given the portal would be international in scope; mechanisms must be in place for stakeholders to select in which language they would prefer to register their study in.

Of the three models identified, this is the most economically demanding. Nevertheless, having pooled metadata records for observational studies accessible through a public facing and searchable catalogue is a novel contribution to epidemiological and public health research. Based on the qualitative results of the survey, having a public health portal specifically for observational studies appeared to be most favoured by participants. A grant proposal detailing how I will take this work forward and build the portal can be found in appendix D.

3.6 Summary of major findings

3.6.1 Review

Most of the studies identified were observational for which there is currently no known mandatory registration process. Results showed that PDF was the most common format in which to provide study protocols and there was limited use of online data visualisation tools. All the studies incorporated social media/other forms of communication in their approach to facilitating data discovery. Therefore, increased investigation is needed into mechanisms to enhance the discoverability of research data in epidemiology and public health.

3.6.2 Online stakeholder survey

The most common form of data was ‘survey’ (27%); the least common was ‘imaging’ (3%). Results show that 82% of respondents thought that the ‘research study question’ should be most easily searchable aspect of a research study. Results also show the terminology most survey respondents were familiar with was the International Classification of Disease (32%).

Most commonly used metadata standard was the Data Documentation Initiative. Survey results show that challenges associated with creating and/or using data documentation can be categorised into 2 groups: a) standardisation; and b) resource availability.

Most popular repository was ClinicalTrials.gov; the least popular was Dryad. The following shows where the median was for aspects important to discoverable data include: a) be on the web – extremely; b) be provided in a machine readable format – extremely; c) be provided in a non-proprietary format – fairly; d) conform with recognised data management standards – extremely; and e) be linked to an underlying conceptual framework or ontology – fairly.

The most common way to hear about data publications was through a colleague. Survey results also show that 22% of respondents who submitted data indicated that data citation was important as it made it easier for readers to locate data. The most popular level of granularity was ‘dataset collections’ - 36% of those who submitted data. Furthermore, 29% of those who submitted data indicated that revisions to longitudinal and regularly changing datasets should be published at regular intervals. Results of the survey also show the challenges associated with the adoption of data publications include: a) significance; b) resource availability (time and cost) and c) research culture.

Areas of importance to data discovery include: a) identification of commonalities and links between studies; b) metadata markup and producing other associated documentation; c) use of technology; and d) consent and other ethical issues. Moreover, results showed that areas of improvement and immediate priority include: a) registration of studies; b) standards (use, metadata standards, quality assessment); c) publications and citations; and consent and other ethical issues

3.6.3 Mechanisms and evaluation

Three mechanisms were identified to enhance research data discoverability: a) data publications; b) application of SWTs; and c) development of a public health portal containing the metadata records for observational studies globally. The public health portal proved most popular with stakeholders.

3.7 Chapter summary

The aim of this chapter was to present the data discoverability study which was composed of a review of a random sample of public health and epidemiology studies and organisations followed by a survey. In this chapter I examined existing approaches to data discovery, identified current challenges and areas of potential improvement, and proposed mechanisms through which funders could enhance the discoverability of their data. The three mechanisms evaluated were: a) data publications; b) linked data on the World Wide Web; and c) public health portal. Following the publishing of the report (Castillo, Gregory et al. 2014), the Wellcome Trust with the Public Health Research Data Forum has begun preparations to take this work forward.

Chapter 4 Research case study 2: Improving metadata quality assessment in public health and epidemiological research

4.1 Introduction

The increase in generation of, and access to, public health and epidemiology data necessitates robust mechanisms to produce metadata. The use of clinical data sourced from electronic health records can greatly benefit from access to metadata and other such documentation (Pathak, Bailey et al. 2013). Researchers can utilise metadata artefacts such as data dictionaries which describe the data to support them in maximising the research potential of the clinical data.

However, in most cases, the extent and quality of available metadata is variable. Some research studies used data publications as a means of providing additional metadata such as the Avon Longitudinal Study of Parents and Children and ELFE, Growing up in France. Whilst others, such as MIDUS Midlife in the US and the Scottish Longitudinal Study used data publications plus provided access to data dictionaries.

Case study: SABE – Survey on Health, Well-being, and Aging in Latin America and the Caribbean

Project SABE ran between 1999 and 2000 and focused on investigating health conditions in those aged 60 and over in Latin America and the Caribbean.[1]

A wealth of information is available on this study from the National Archive of Computerized Data on Aging, but there is currently no mechanism to assess the quality of these metadata. Consequently, stakeholders wanting to assess the metadata available, and potentially compare the metadata to previous versions, are unable to do so. Furthermore, there appears to be no mechanism to monitor changes to the metadata from the user's perspective.

The aim of this research case study is to create and evaluate a novel metadata quality assessment framework. The novel framework will address the current lack of assessment criteria to evaluate metadata in public health and epidemiological research settings.

1. http://www.icpsr.umich.edu/icpsrweb/NACDA/studies/3546

Having access to good quality, wide-ranging metadata across the stages of the research data lifecycle is critical to enabling stakeholders to manage their research data more effectively and to investigate the causes of disease at a greater depth. Research artefacts such as data publications can describe the study protocol and the data dictionary the variables; combined these can provide researchers with the contextual information needed to enable them to analyse the research data more effectively.

This chapter presents the metadata quality study in which I suggest a novel, quality assessment framework for biomedical metadata. This framework will be a word processed document which can be used alongside your metadata management software and is not restricted to one particular operating system. This work addresses this gap in knowledge relating to metadata management within the context of epidemiological and public health research.

4.2 Study aim and objectives

The aim of this work was to create and evaluate a novel metadata quality assessment framework to support the administration of metadata and improve overall quality.

This study had four objectives: a) describe the current state of the art in metadata quality assessment of biomedical research data through a systematic literature review; b) identify key metadata quality assessment dimensions of biomedical research data; c) create and evaluate a novel framework for assessing metadata quality for epidemiology/public health research data; and d) apply the framework to test cases and perform stakeholder interview by way of evaluation.

4.3 Methods

4.3.1 Literature review

I performed the literature review using the PRISMA checklist(Moher, Liberati et al. 2009) for guidance. This systematic literature review aimed to answer the following questions:

1. What is the current state of art in metadata quality assessment?
2. Which methods of metadata quality assessment are available?
3. Are there any methods specifically designed for use in public health and epidemiological research settings?

4.3.1.1 Eligibility criteria

To be included in the review, the publication had to be available in English and accessible through open access or using login credentials. The publication had to clearly define a method of quality assessment and ideally provide examples of where the method has been applied.

4.3.1.2 Information sources and search terms

The literature review was performed in July 2014 using cross-disciplinary databases to identify literature for inclusion in the review. The databases were, ACM Digital Library, BioMed Central, CINAHL Plus, Cochrane Library, EMBASE, Lecture Notes in Computer Science, JSTOR, PubMed, SCOPUS and Web of Science. I included both biomedical and computer science databases to help me identify literature which may have been indexed under research domains such as librarianship or archive management. Additional sources included Google and Google Scholar, and forward citation tracking (Kuper, Nicholson et al. 2006) was used to help source any other potential methods of quality assessment. The searches included all forms of literature and were not restricted in terms of publication date or location, Supplementary Table 3.

I used the following search terms: ‘epidemiology’, ‘metadata’, ‘metadata quality assessment’, ‘metadata quality dimensions’, ‘metadata quality evaluation’, ‘public health’, ‘public health and epidemiology’, ‘quality assessment’, ‘quality evaluation’.

4.3.1.3 Study selection

Once the publications were identified, I firstly removed duplicates. I then screened each publication by reading the title and abstract. Based on these, I then removed publications if they were irrelevant, not available in English or if the full text was unavailable. The remaining publications were then fully reviewed and I removed an additional set of manuscripts due to their ineligibility.

The publications included in the review had the following recorded: a) title; b) aim; c) conclusion; d) method of quality assessment – a brief description of the approach to quality evaluation; and e) identified metadata quality dimensions – identified underlying principles of quality. As each approach is recorded, it was assigned a unique ID.

4.3.2 Online stakeholder survey

To describe the design and development of the online stakeholder survey in a systematic way, I have chosen to follow the CHERRIES(Eysenbach 2004) statement as adopted by the Journal of Medical Internet Research.

4.3.2.1 Design

The survey has five sections: demographics, metadata, tools and technologies, metadata usability, and quality assessment. I designed and developed my survey using REDCap (Research Electronic Data Capture) version 5.7.5.(Harris, Taylor et al. 2009) which is a secure, online data collection tool. A copy of the survey can be found in the appendix B.

In the metadata section, the types of metadata listed were identified through desk research (Pollock and Hodgson 2004; Zeng and Qin 2008; Miller 2011). In the tools and technologies section, the list of metadata standards was partially based on a list provided by the Digital Curation Centre on their website(Digital Curation Centre 2014) and the outcomes of the discoverability study. In the usability of metadata section, aspects of potential importance are based on the exploration of metadata quality impacting the different stages of the research data lifecycle from chapter one, 1.4.2. The aspects are also based on the findings from chapter two. In the quality assessment section, the suggested quality dimensions are based on the outcomes of the data discoverability study(Castillo, Gregory et al. 2014) and results of the review.

4.3.2.2 Ethical approval and informed consent process

Ethical approval was not required and implied consent to partake in the study was assumed from the individual through their decision to submit data using the surveys. All data collected was anonymous and contact details were provided should (potential) participants wish to contact me for further information or clarification. Given the anonymity of the survey participants, participants were asked if they would like to potentially partake in any other aspects of the study. Therefore, participants were asked to provide contact details if they choose to reveal their identity and give consent to be contacted. Any identities and contact details provided are and will remain confidential.

4.3.2.3 Recruitment process and description of the sample having access to the questionnaire

The following mailing lists were used to circulate the survey:

- UCL data management – public health: staff and students involved in public health data management at UCL
- UCL Centre for Health Informatics and Multidisciplinary Education (CHIME) - staff and students affiliated with UCL CHIME
- DDI users mailing list – all those interested in and/or using the DDI metadata standard
- UCL Epidemiology and public health student mailing list
- UCL Epidemiology and public health staff mailing list
- JISC Research data management - all research domains
- JISC Managing research data – all research domains
- JISC Public health mailing list

Promotional material was circulated at the following conferences to raise awareness of the study: American Medical Informatics Association 2014 Annual Symposium (Washington DC, USA), and EDDI 2014 – Annual European DDI User Conference (London, UK)

In addition to these, representatives of the signatories and supporting organisations of the Public Health Research Data forum were also contacted and asked to circulate the surveys. I also sent an invitation to complete the survey to a colleague at the Nature Publishing Group and requested the invitation be shared with their colleagues too. Calculating response rates was unfeasible given the approach adopted to recruit participants to the survey.

4.3.2.4 Survey administration

The invitational email included a short description of the metadata quality, study, a link to the survey, mention and hyperlink to the data discoverability report(Castillo, Gregory et al. 2014) and the request that participants forward the invitation to any parties they feel may be interested.

The metadata quality survey ran from November 2014 to February 2015. A set of reminder emails were sent two weeks before the survey was due to close. The survey begins with a short introductory paragraph to the study describing the aims and objectives of the survey and the length of time it should take to complete (10 minutes).

The survey questions (all optional) were collated according to theme and spread across different pages to ease the process of stakeholders submitting data (Schleyer and Forrest 2000). Boolean logic was incorporated to customise the survey based on previously submitted answers by automatically adding or removing subsequent questions (Wyatt 2000). In doing so redundant questions were removed from the survey and only those relevant to the respondent were presented on screen. I also included open ended questions to help facilitate the capture of qualitative data. (MacKenzie, Wyatt et al. 2012) I also included a Likert scale (5 points) to help the stakeholders convey their opinions. The options were as follows: ‘not at all’, ‘slightly’, fairly’, ‘extremely’, and ‘essential’.

4.3.3 Framework definition and evaluation

I analysed the results using a Grounded Theory (Glaser and Strauss 1967) approach which enabled me to develop hypotheses based on the data helping to produce a truer reflection of current events. I used a hermeneutic approach when inductively and iteratively collating themes (Dahlberg 2010; Svanstrom, Andersson et al. 2016). I analysed the quantitative data using SPSS. The metadata quality assessment framework consisted of four sections: a) general information; b) tools and technologies; c) usability; and d) management and curation. I identified the four sections by initially grouping together the suggested headings according to the issues they addressed. Once all the suggested headings had been grouped, I assigned a single more generalised heading. For example, the ‘indexing in catalogues’, ‘encoding and exchange standards’ headings are grouped under ‘Tools and technologies’. The requirements of the framework were based on findings from the systematic literature review and the online stakeholder survey.

The evaluation of the framework is a two stage process. Part A involved iteratively applying and evaluating the framework to three test cases. After each application, the framework was improved. The first test case was the Millennium Cohort Study (Supplementary Table 4). The second and third test cases were: Midlife in the United States (MIDUS) (Supplementary Table 5) and the Danish National Birth Cohort (DNBC) (Supplementary Table 6). Part B involved engaging with stakeholders in public health and epidemiological research to further evaluate the framework and provide insights to facilitate implementation of the framework into research policy and practice. I engaged with stakeholders through a combination of face-to-face discussions and email conversations.

4.4 Results

4.4.1 Literature review

4.4.1.1 Study selection

A total of 11 publications were identified and included in the review. Figure 4-1 shows how the final set of publications was identified and Table 4-1 shows results of the review.

Figure 4-1 Metadata quality literature PRISMA flow diagram

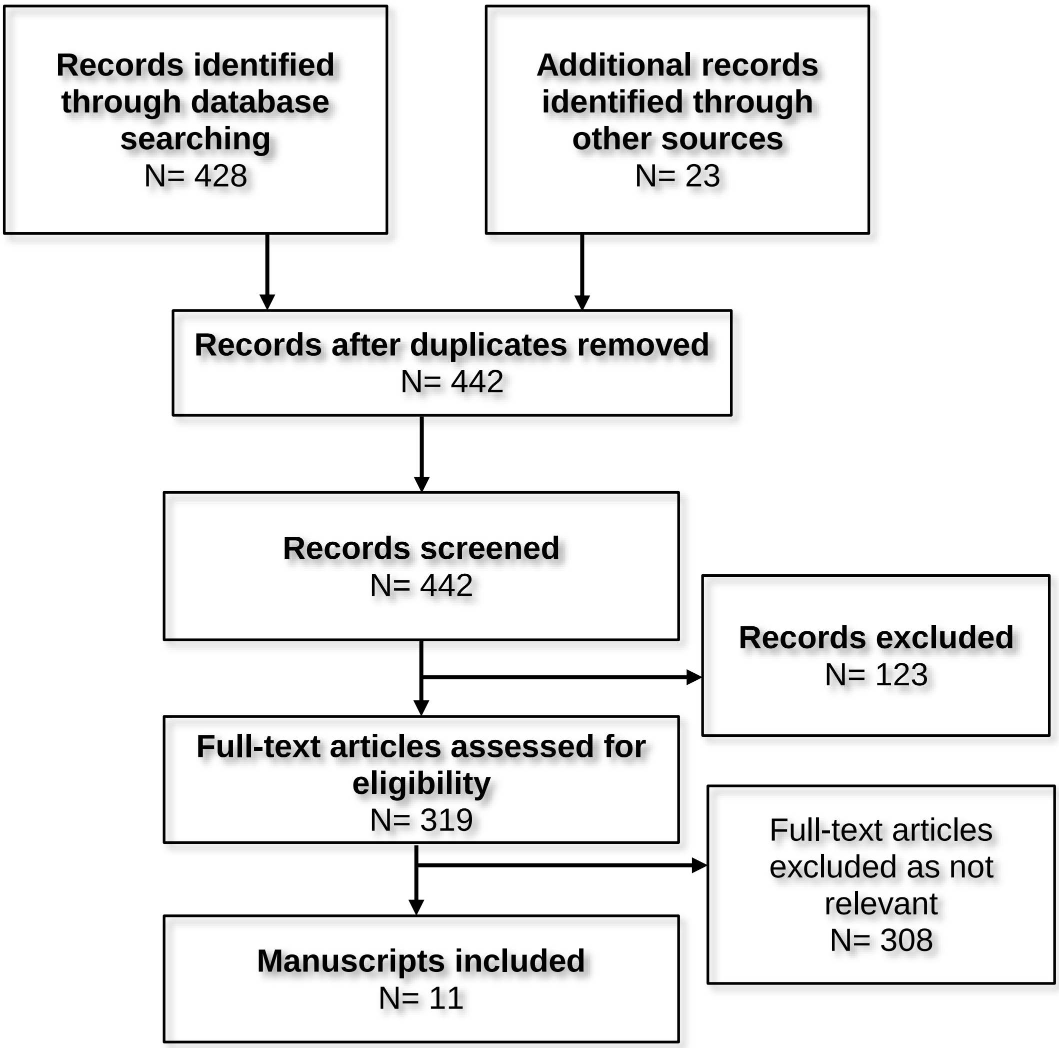

Records identified through database searching
N= 428
Additional records identified through other sources
N= 23
Records after duplicates removed
N= 442
Records screened
N= 442
Records excluded
N= 123
Full-text articles assessed for eligibility
N= 319
Full-text articles excluded as not relevant
N= 308
Manuscripts included
N= 11

Table 4-1 Review of methods of metadata quality

ID	Title	Aim	Conclusion	Method of quality assessment	Identified metadata quality dimensions
1	Assessing metadata quality: findings and methodological considerations from an evaluation of the US Government Information Locator Service (GILS) Moen, Stewart et al. (1998)	To investigate the use of qualitative and quantitative techniques to assess metadata quality. The objectives are: • establish how accurate the GILS records are • compare completeness of records • identify common characteristics of GILS records • assessment serviceability of records	The identification of 4 dimensions of quality operationalized by the indicators successfully enabled the GILS records to be evaluated. It was also noted that development of the records needs participation of all those involved in GILS records. Findings will inform development of records and improvement of service.	17 indicators divided between 4 categories	Completeness, Accuracy*, Accessibility
2	The continuum of metadata quality: Defining, expressing, exploiting Bruce and Hillmann (2004)	To report on: • quality dimensions • levels of metadata quality • suggest potential indicators of quality • short and longer term quality	A total of 7 dimensions of quality were suggested accompanied by criteria and indicators. The inability to exhaustively list quality indicators is acknowledged as is the need for further development.	7 indicators	Completeness, Provenance, Timeliness, Accuracy, Accessibility, Logical consistency and coherence, Conformance to expectations
3	Metadata quality evaluation: Experience from the	To develop and implement a framework to assess metadata quality within the Open	Successful development of a scalable method to evaluate metadata quality.	7 metrics	Accuracy, Completeness

ID	Title	Aim	Conclusion	Method of quality assessment	Identified metadata quality dimensions
	open language archives community Hughes (2005)	Languages Archives Community (OLAC) as part of the Open Archives Initiative (OAI).	Acknowledgement is given to the lack of support for qualitative assessment and the vision for the wider OLAC community to be assisted by the proposed framework.		
4	An assessment of metadata quality: A case study of the National Science Digital Library Metadata Repository Bui and Park (2006)	To report on: • Metadata quality of records in NSDL repository • Extraction of Dublin Core metadata • Present preliminary results	A total of 1311169 metadata records were extracted using OAI-PMH 2 protocol. The most important DC elements were (random order): descriptor, subject, title, identifier, type and creator.	15 Dublin Core elements + other local elements	Accuracy, Completeness, Consistency, Frequency
5	Author-generated Dublin Core metadata for web resources: A baseline study in an organisation Greenberg, Pattuelli et al. (2006)	To investigate the Dublin Core metadata standard may be used to produce quality metadata for resources on the national Institute of Environmental Health Sciences web site. More specifically, it is to evaluate manually-generated metadata.	In using the Dublin core elements, the authors were able to determine users' opinions on metadata and the user interface. Results show users were able to produce metadata in accordance with the NIEHS-Dublin Core schema. A limitation of the study is the sample size due to the nature of the study – more analysis is needed.	15 Dublin Core elements	Accuracy, Completeness, Timeliness, Interoperability, Accessibility

ID	Title	Aim	Conclusion	Method of quality assessment	Identified metadata quality dimensions
6	A framework for information quality assessment Stvilia, Gasser et al. (2007)	To present a domain-independent quality framework. Such a framework can be used a basis for the development of more specialised methods of assessments.	Development of quality taxonomy enables problems associated with quality to be studied in greater depth. The generality of the framework lends itself well to further development.	22 dimensions categorised according to structure	Totality, Accuracy, Accessibility, Interoperability
7	Toward releasing the metadata bottleneck A baseline evaluation of contributor-supplied metadata Wilson (2007)	To assess the contributions made by persons other than metadata experts to the assurance of quality. The study focuses on ascertaining completeness of records, establishing the types of errors and identifying additional metadata to the records.	This study uses the abstracts from RILM (music literature) and serves as a basis for future work on metadata quality. The potential of contributors as a source of additional metadata needs to be further explored as do opportunities to develop systems to ingest this type of metadata to improve overall quality.	12 – Record analysis 8 – Abstract content analysis	Totality**, Accuracy**, Completeness**
8	A Conceptual Framework for Metadata Quality Assessment Margaritopoulos, Margaritopoulos et al. (2008)	To develop and present a conceptual framework (using a court of law analogy) for evaluating metadata quality using logic rules.	Application of the framework took into consideration structural and semantic relationships to evaluate quality. Future work includes that of deriving metrics to quantify metadata quality.	3 categories of logic rules: rules of inclusion, rules of imposition and rules of restriction	Completeness, Correctness
9	Automatically characterizing resource quality for	To report on: • previous and related research	Presentation of 7 low level quality indicators accompanied by computational methods to	12 (general descriptors) 7 (low level	Totality, Accessibility, Conformance,

ID	Title	Aim	Conclusion	Method of quality assessment	Identified metadata quality dimensions
	education digital libraries Bethard, Wetzer et al. (2009)	• approaches to study • identify general areas of quality and then develop low level metrics • accuracy of machine learning techniques	automatically assess quality as part of educational libraries. Using machine learning techniques, they were able to achieve over 80% accuracy.	indicators)	Interoperability
10	Automatic evaluation of digital libraries with 5Squal Moreira, Gonçalves et al. (2009)	Development and evaluation of the tool, 5Squal, to enable automatic quantitative evaluation. The tool is based on the model proposed by Gonçalves, Moreira et al. (2007)	Following evaluation, the tool meets the need for a method of evaluation of digital libraries. Users liked the user interface and graphs. Possible areas of future work include inclusion of additional dimensions/metrics and standards.	8 (2 relate specifically to metadata)	Timeliness, Completeness, Conformance, Accessibility
11	Automatic evaluation of metadata quality in digital repositories Ochoa and Duval (2009)	To present previous work on approaches to quality evaluation using the framework developed by Bruce and Hillmann (2004), as a basis for metadata quality metrics specifically for information completely manually, automatically or both.	The metrics were evaluated in three ways: 1) comparison of metrics and human reviews, 2) looking at metadata sets and 3) metrics as low-quality filters. Authors suggest the metrics be used as a baseline for comparison with other new metrics.	7 – based on framework proposed by Bruce and Hillmann (2004)	Completeness, Provenance, Timeliness, Accuracy, Accessibility, Logical consistency and coherence, Conformance to expectations Provenance

*In terms of formatting only; ** Using record level dimensions only.

See Table 4-2 **for metadata quality definitions.**

4.4.1.2 Synthesis of results

Results of the review showed that the greatest total number of metrics used as part of a single approach was 22 (Stvilia, Gasser et al. 2007). I also identified that several methods group the metrics according to different criteria (Moen, Stewart et al. 1998; Stvilia, Gasser et al. 2007; Wilson 2007; Moreira, Gonçalves et al. 2009). This review also identified a method which uses a series of logic rules to divide the assessment. These were: a) rules of inclusion; b) rules of imposition; and c) rules of restriction (Margaritopoulos, Margaritopoulos et al. 2008).

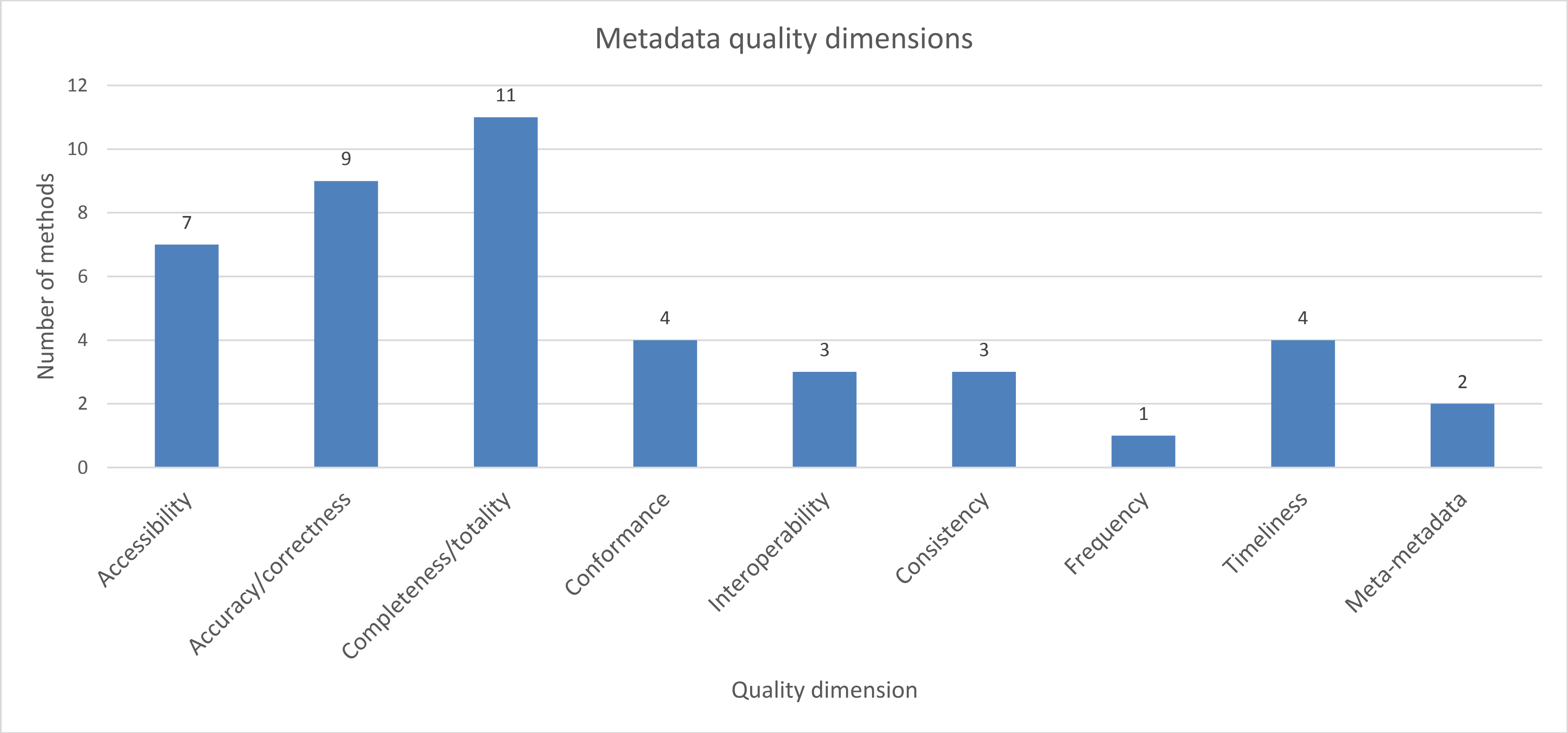
Collectively, the methods detailed in this manuscript presented a total of nine different dimensions of quality. These dimensions of quality were measured using the metrics identified in each method. Following the review, a number of common metadata qualities emerged. The amalgamated set of metadata quality dimensions with definitions can be found in Table 4-2.

Table 4-2 Dimensions of metadata quality and associated studies

Identified dimension of metadata quality	**Study ID**
Accessibility - Extent to which the metadata can be accessed	1, 2, 5, 6, 9, 10 and 11
Accuracy/correctness - Correctness of the metadata	1 - 8, 11
Completeness/totality - Presence of all metadata	1-11
Conformance - How well the metadata conforms to expected standards	2, 9 - 11
Interoperability - Extent to which metadata can be exchanged and used without problems)	5, 6, 9,
Consistency - Does the metadata form, content etc. change throughout the document or does they remain	2, 4, 11
Frequency - How often a metadata element is used	4
Timeliness - How current the metadata are	2, 5, 11
Meta-metadata - Metadata about the metadata	2, 11

The most commonly occurring dimension was completeness/totality with all the approaches incorporating this dimension; the least common dimension was frequency(Bui and Park 2006). Figure 4-2 shows a comparison of the quality dimensions across the 11 manuscripts included in the review. Use of metadata standards were also identified as part of the review. Here, the simple Dublin Core elements had been incorporated into two methods (Bui and Park 2006; Greenberg, Pattuelli et al. 2006).

Figure 4-2 Metadata quality dimensions



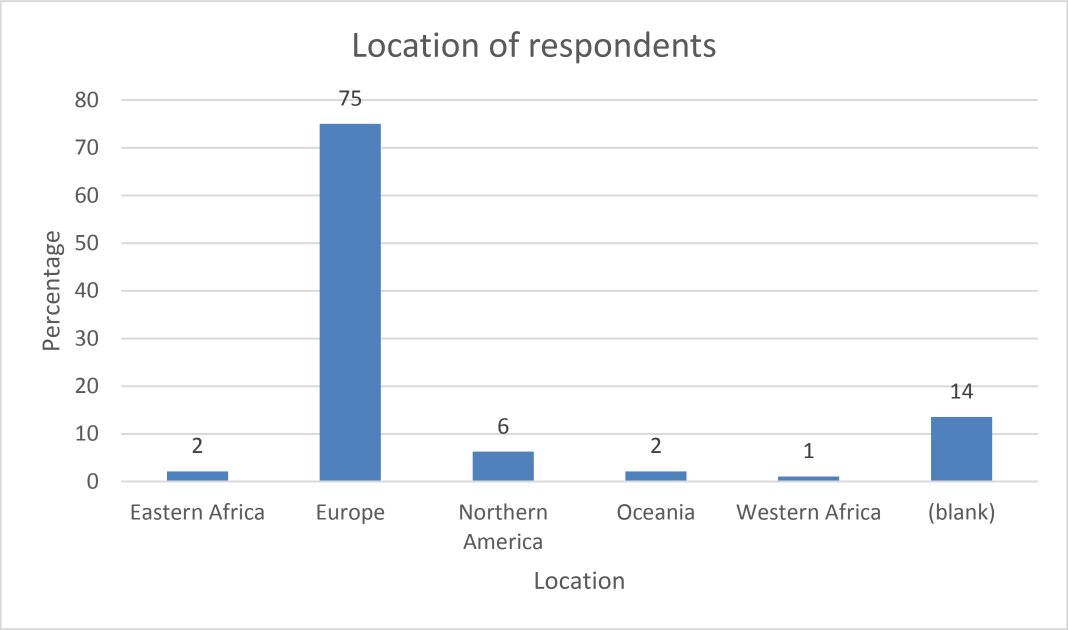
4.4.2 Online stakeholder survey

The following are the results of the online stakeholder survey. Where quotations have been used, these have been copied verbatim; in places words have been added in parentheses to aid understanding.

4.4.2.1 Participant demographics

Ninety-six individuals submitted data using the survey; most of whom indicating they were employed by a university and currently located in Europe Figure 4-3. The most commonly indicated role in public health and epidemiology (multiple selection enabled) was ‘Data user’.

Figure 4-3 Location of respondents



4.4.2.2 Section one of the survey results: Metadata

The participants were asked to indicate how often they used metadata. A total of 47 people answered this question with the majority indicating they had used metadata on a regular basis (Table 4-3).

Table 4-3 Use of metadata

	Frequency	Percent
Never	6	6%
Sometimes	6	6%
Regularly	15	16%
Frequently	9	9%
Very frequently	11	11%
Total	47	49%

The participants were also asked to indicate which types of metadata they used. The most commonly selected type of metadata was descriptive with a total of 37 votes followed by administrative with 30 (Table 4-4). Of those that indicated 'other', the issues of it not always being clear which type of metadata was being used, and developing metadata for future use was raised.

Table 4-4 Types of metadata

	Responses		
	N	Percent	CI
Administrative	30	31%	22-41
Descriptive	37	39%	29-49
Microdata	15	16%	9-24
Semantic	12	13%	6-21
Other	2	2%	0-7
Total	96	100.0%	

Participants were asked to indicate in which of the specified formats the metadata they routinely handled had appeared. Results indicate, of those that answered this question the most commonly handled format of metadata was PDF(s) with 27; the least commonly indicated was RDF.(Table 4-5) Of those which indicated other, additional responses included: SAS files, Stata syntax files, SQL and txt.

Table 4-5 Formats of routinely used metadata

	Responses	
	Number	Percent
PDF(s)	27	23%
Spreadsheet(s)	18	15%
Word© document(s)	20	17%
XML	25	21%
RDF	6	5%
HTML	17	14%
Other	6	5%

The respondents were asked to indicate in which stages of the research data lifecycle they handled metadata. Again, multiple selections were enabled. Results show the stage at which most respondents used metadata was the analysis stage; the stage which received the fewest votes was data destruction. (Table 4-6)

Table 4-6 Handling metadata across the research data lifecycle

Handling metadata across the research data lifecycle

		Stages of the research data lifecycle										Total
		Conceptualisation	Creation or receipt	Appraisal & Selection	Analysis	Preservation action	Access, use and reuse	Transformation	Data destruction	Archive management	Administration	
Role in public health[a]	Archivist / librarian	4	4	3	1	3	4	2	0	4	3	6
	Clinician / clinical advisor	3	4	3	6	1	5	3	0	3	2	8
	Data provider	8	11	7	10	10	12	8	2	4	5	17
	Data user	10	12	11	16	7	12	12	3	3	9	20
	Funding agency	0	1	0	0	0	1	0	0	1	0	1
	Policy maker	1	3	2	2	1	3	1	1	2	2	3
	Observer	1	1	1	1	1	1	1	1	1	1	1
	Other	4	2	2	4	2	4	3	1	3	2	7
Total		18	21	17	23	16	24	18	3	13	14	39

a. Dichotomy group tabulated at value 1.

The granularity of metadata was also addressed with participants being asked to indicate at which levels metadata should be made available. The most popular level was ‘Research study level’ (38 votes) followed by ‘Variable level’ (37). Through the ‘if other, please specify’ option, one respondent noted that metadata should become available each time there is a, “Major transformation e.g. merging”. (Table 4-7)

Table 4-7 Granularity of metadata

	Responses	
	Number	Percent
Research study level	38	28%
Single dataset / sweep of data	35	26%
Variable level	37	28%
Each time a change is made to the data	23	17%
Other	1	1%
Total	134	100.0%

The last question in this section requested participants describe the main barriers to creating and/or using metadata in biomedical research. The barriers can be categorised into the following Table 4-8:

Table 4-8 Main barriers to creating and/or using metadata in public health and epidemiological research

Main barriers to creating and/or using metadata in biomedical research	Findings
a) Lack of skills and experience	This refers to the inadequate training and in particular, "a lack of guidance on how to create metadata for normal data users…many internal initiatives that request te same information without in the end generating sustainable end product, thus wasting the time of metadata creators" Three respondents commented on the lack of awareness relating to metadata use. For example, one respondent said that, "…a lot of people in my team are not aware they are using metadata…it would be nice if the metadata were provided by the data providers theirselves (HSCIC, ONS etc.)" Whilst another commented on a, "lack of understanding of why metadata is important (until it's too late!)" It was also noted by one respondent that a "…big problem is that the process of document the data starts to late. The research doesn't have a structured plan from the beginning that includes documentation of metadata."
b) Inconsistencies	Namely formatting of metadata and the approaches to "…developing and categorising models (and) usable interfaces for development". One respondent in particular commented, "…sometimes I am sent data with no associated metadata and therefore don't know the definitions of the variables…"
c) Inadequate tool availability	Three respondents commented on a lack of tools with one stating, for example, there being a "…lack of tools to create data dictionaries across sweeps…" Another commented on having the relevant support to use the

	tools once these were available. Another related issue is being able to extract metadata from older, potentially inaccessible versions of software.
d) Lack of standards	Five respondents commented on a lack of standards with one respondent stating, "as far as I'm, aware there are no standards so quality can very". Another commented on the, "agreement to adopt common standards".
e) Ethics	Two respondents commented on restrictions put in place due to data protection and other issues related to ethics impacting the creation and/or use of metadata; one respondent in particular comments on "…trust, privacy and security…"
f) Inadequate resources: cost and time	Six respondents commented on a lack of resources to create and curate metadata effectively.

My work can help to address these issues through a number of different ways. Firstly, the framework provides a robust method to assess metadata quality which also serves to help better educate users through the provision of guidance and definitions throughout the framework. Secondly, the framework serves as a tool to help users systematically assess metadata quality in a replicable way. This help to standardise the method of quality assessment. Thirdly, the availability of a framework reduces the time otherwise taken by users to create their own tool to assess metadata quality.

This tool has already been shared with a data manager at another institution and is currently under review.

4.4.2.3 Section two of the survey results: Using tools and technologies

Clinical terminologies and classification systems: This section of the survey focused on tools and technologies. The aim of the initial question was to determine how tools and technologies were selected. The most commonly selected answer was ‘Standard practice’ while the least commonly selected answer was ‘Funder requirement’. (Table 4-9). A total of 4 people indicated ‘other’ with additional suggestions including, “What I am used to”, “Highlighted via mailing list”, “based on requirements of the end user or data provider” and “Do my own research into tool”.

Table 4-9 Selecting tools and technologies

	Responses	
	N	Percent
Funder requirement	9	14%
Suggestion from colleagues	21	33%
Standard practice	29	46%
Other	4	6%
Total	63	100.0%

Survey respondents were then asked about their experiences with encoding standards. Of those which submitted data, the classification system most people had come across was International Classification of Diseases followed by Medical Subject Headings. Both Diagnostic and Statistical Manual of Mental Disorders and Logical Observation Identifiers Names and Codes each received 0 votes (Table 4-10). Of those which provided additional answers, other clinical terminologies included “Multilex”.

Table 4-10 Clinical terminologies

Clinical terminologies

	Role							
	Archivist / librarian	Clinician / clinical advisor	Data provider	Data user	Funding agency	Policy maker	Observer	Other
ICD	1	4	6	14	1	2	1	5
MeSH	1	4	3	8	1	2	1	5
OPCS	0	1	4	6	0	0	0	1
Read Codes	0	3	2	11	0	1	1	3
SNOMED CT	1	3	4	6	1	2	1	3
Other	0	0	3	2	0	0	0	0
Total	1	6	8	16	1	2	1	7

Respondents were also given the opportunity to describe any difficulties they faced when using clinical terminologies. Results can be categorised into the following:

A lack of medical knowledge: respondents felt that due to a lack of medical knowledge, the time taken to understand the terminology, and hence utilise the clinical information, increased. These findings suggest greater explanatory information is needed to support the use of clinical data for research purposes.

Lack of ease of use: one respondent commented on the, "sheer volume of terms" whilst two others commented on "translating from one coding system to another…" and "overlapping codes Inappropriate codes Missing codes". A third commented on the appropriateness of using certain coding systems for research purposes, for example, "MeSH not adequate for public health research". While another respondent commented on the "variable coding by clinicians" as being another problem associated with use of clinical terminologies. Another two respondents commented on the structure and complexity of the codes themselves. One wrote, "…codes used not the best for the situation, some confusion over code hierarchies". While the other wrote,

> "the terminologies (and the codes) are complex and with many levels. Not a 100pct. uniq connection (ICD) between the codes and terminology over time"

These findings suggest that the size, complexity and application of clinical terminologies and classification systems can negatively affect how easily they may be utilised in research. Consequently, researchers are experiencing difficulties in harnessing these controlled vocabularies to support the realising of the potential research benefits of clinical data.

Variable access to classification system - one respondent commented that, "…the WHO classification is easily available online, but the HES classification is harder to obtain…". which negatively impacts the speed at which codes may be verified. These findings suggest that the inconsistent access to classification systems can negatively impact their application in research. Consequently, the respondents reported having to spend in increased amount of time trying to verify encoded clinical information.

Metadata standards: The most commonly used metadata standard by the respondents was the Data Documentation Initiative 2(Codebook) followed by Data Documentation Initiative 3 (Lifecycle). The least commonly used metadata standards with no votes each were, Minimum Information for Biological and Biomedical Investigations (MIBBI), Observ-OM and Open Microscopy Environment eXensible Markup Language (OME-XML). (Supplementary Table 2) Additional metadata standards include Metadata Object Description Schema (MODS), “EHR standards – openEHR, HL7 and EN13606”.

A total of 11 people indicated that they used metadata catalogues to help improve the discoverability of their research. Of those which submitted data, challenges associated with catalogues to improve discoverability include the catalogue not being fit for use, and having researchers and other stakeholders use/contribute to the catalogue. For example, one respondent said there was an,

> “…impossibility to use existing tools as they don’t apply to longitudinal studies with several data sweeps.”

This was echoed by other respondents who commented that a challenge was, “having people contribute their models!” and that they are “normally treated by researchers as add on, thus a cost with no benefit to their work”.

With regards to using catalogues to identify and characterise metadata, a total of 16 people indicated they had used a catalogue in this way with six stating they had experienced problems. Problems included not having the granularity required to assess the data, and the method of selecting variables being inconvenient,

> “Data were described sweep by sweep instead of giving an overview across sweeps. Using 'shopping baskets' to select variables is extremely inconvenient and it is very slow when hundreds of variables are needed”

Other problems included limited knowledge of standards such as “DDI” and being able to maintain the catalogue itself.

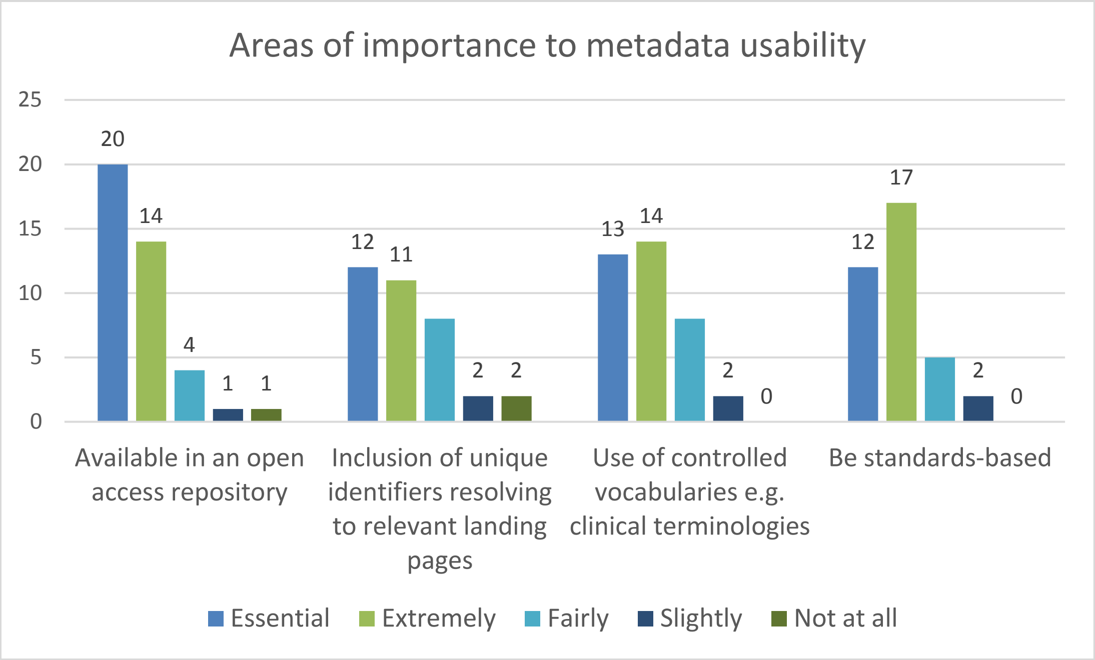
Survey participants were asked to indicate whether they had used SWTs when handling metadata. A total of five people indicated they had with two sharing their experiences. Challenges shared include, “insufficient resources” and there being a “small community” with a “higher barrier of knowledge”.

The survey then asked the respondents to indicate whether they had used any kind of metamodel as part of their research. A total of five people said they had used some kind of metamodel with examples including, “HL-7”, “OpenEHR”, “EN ISO 13606”.

4.4.2.4 Section three of the survey results: Metadata usability

In order to identify areas of importance to metadata usability, respondents were asked to share their opinions of different aspects of usability. Results show that the majority view the metadata being available in an open access repository as being essential whilst 17 people view being standards-based as extremely important to usability. A total of two people thought that the inclusion of unique identifiers resolving to relevant landing pages as being not at all important to the usability of metadata (Figure 4-4).

Figure 4-4 Areas of importance to metadata usability



Respondents were also given the option to share additional suggestions of aspects important to metadata usability. Additional responses include, the metadata being "…easy to create and edit by non-programmers, using user-friendly tools" and for there to be an "encouragement of common understanding between contributors and users". These findings suggest that more is needed in the way of tools to support the creating and managing of usable metadata. These findings also suggest more is needed in the way of harmonised definitions of terms to enable multidisciplinary teams, including contributors and user, to work even more closely together. Problems associated with usability include having the training and documentation to use the metadata in particular having, "…a human readable as well as machine readable…" formats of the metadata. Other issues raised include using open source technologies, the process of "doi minting" and having access to, "…semantic mapping to relevant terms."

I fed these issues relating to usability into the framework in a number of different ways. For example, in the Tools and Technologies section, I suggested users consider whether the metadata have been included in catalogues and if any use had been made of encoding standards – controlled vocabularies. Further to this, I also suggested users consider whether other standards used such as exchange standard and metadata standards. With regards to inclusion of unique identifiers, in the framework I suggested, in the General section, that users included location of the research artefacts and links resolving to studies, sweeps or publications.

By having access to metadata in both machine and human readable formats, researchers can have an increased flexibility in terms of processing the metadata - this can either be done manually using the human-readable format or electronically using the machine-readable format.

4.4.2.5 Section four of the survey results: Quality assessment

To determine the dimensions of quality important to public health and epidemiological research metadata, the participants were asked to vote (multiple selection enabled) for dimensions from a pre-defined list (Table 4-11). Results show of those who submitted data, accuracy was viewed as being most important to epidemiological and public health metadata. The second and third most important were, accessibility and discoverability respectively. The dimension participants viewed as being least important was meta-metadata. Of those which provided additional information, suggestions for dimensions were not provided; instead the following two comments, “before meta metadata let’s get metadata clear and well understood” and the second, stating that other research communities use metadata too, were made.

Table 4-11 Dimensions of metadata quality

		Responses	
		N	Percent
Dimensions of metadata quality[a]	Accessibility (extent to which the metadata can be accessed)	34	14.5%
	Accuracy (correctness of the metadata)	35	14.9%
	Appropriateness (extent to which the metadata are relevant)	24	10.2%
	Comprehensiveness (extent to which the metadata are complete)	24	10.2%
	Discoverability (how visible the metadata are - can it be easily found)	33	14.0%
	Extendibility (extent to which the metadata may be easily extended)	14	6.0%
	Interoperability (extent to which metadata can be exchanged and used without problems)	21	8.9%
	Meta-metadata (metadata about the metadata)	10	4.3%
	Timeliness (is the metadata current, inclusion of temporal information)	20	8.5%
	Versionability (extent to which a new version may be easily created)	18	7.7%
	Other	2	0.9%
Total		235	100.0%

a. Dichotomy group tabulated at value 1.

Results also showed that most of the respondents, who submitted data sometimes assessed its quality while eight people never assess quality (Table 4-12).

The survey then asked participants if they used any kind of formal metadata assessment criteria. A total of 28 people answered this question of which only one said they used some kind of criteria. This was a, “built in diagnostics within metadata editor”.

Table 4-12 Frequency of metadata quality assessment

	Frequency	Percent
Never	8	8%
Sometimes	19	20%
Regularly	6	6%
Frequently	3	3%
Very frequently	1	1%
Total	37	39%
System	59	61%
Total	96	100%

Challenges associated with assessing metadata quality in epidemiological and public health can be categorised into technical and cultural. For example, several participants commented on the lack of guidance to assist quality assessment with one respondent in particular commenting that one of the problems is identifying, “…the best way to determine quality” while another wrote, “I haven’t thought about this”. This lack of guidance and awareness of metadata quality assessment demonstrates the current unmet need of a formalised method of assessing metadata quality within the public health and epidemiological research context. Having access to a formalised method of metadata quality can potentially help researchers to improve the way in which metadata are managed.

Another perceived challenge was a lack of domain-specific knowledge that negatively impacting how well respondents understood the metadata.

For example, one respondent commented on a “lack of experience and knowledge of the area” whilst another commented on the “lack of knowledge of biomedical analysis”. This reported lack of domain-specific knowledge demonstrates the needed for increased support, possibly through improved training and access to formalised guidelines, to enable researchers to better understand and utilise metadata.

Respondents also reported how the limited availability of tools and resources negatively impacted how well they were able to assess metadata quality. For example, respondents reported not knowing about software and other such tools and secondly not being able to access these. They also reported on how time-consuming a process like metadata quality assessment is and how finding this time was problematic. These findings demonstrate the need for a formalised method of assessing metadata quality which can be implemented relatively quickly and without disruption to daily work routines. These findings are also indicative of the lack of focus on the issue of metadata quality to date and demonstrates the need for greater emphasis on not only assuring metadata quality but also the potential negative impact of not doing so.

4.4.3 Framework definition

The following describes the sections of the framework:

4.4.3.1 Novel metadata quality models

In the ‘General information’ model, the completeness refers to how comprehensive the metadata are, and the granularity looks at the level at which the metadata available, e.g. research study level. The types of metadata, includes administrative, descriptive, microdata and semantic. The formats includes but are not limited to, PDF, Spreadsheet, Word processed document, XML, RDF and HTML. In having access to metadata in different formats, the discoverability and accessibility of research data can be potentially enhanced. Metadata provided in RDF for example, can be published on the World Wide Web whilst its characteristic scalability can be harnessed facilitating the integration of multiple disparate resources in support of clinical research (Pathak, Kiefer et al. 2012a; Pathak, Kiefer et al.

2012c). This also helps to improve accessibility as the metadata can be retrieved in a number of different ways; there is also scope for inclusion in catalogues through use of metadata harvesting tools – both of which contribute to the enhanced discoverability of research data.

In the 'Tools and technologies' model, accuracy refers to the use of clinical terminologies and in particular the provision of any codes assigned to pre-determined answers and the categories these codes may fall into; more specifically, it suggests stakeholders look at the way in which clinical information were encoded using clinical terminologies. Structure refers to how the metadata are presented; for example, some studies provide tables containing sweep-level metadata with downloadable files inclusive of descriptions of any use of clinical terminologies. Provision of clearly structured and unambiguous metadata can help stakeholders to navigate through the metadata more easily. Accessibility can also be enhanced through inclusion of the metadata in a public facing, searchable catalogue such as the CALIBER portal (Denaxas, George et al. 2012). Use of portals or catalogues can also help to enhance the discoverability of research data by providing collections of metadata from a range of epidemiological and public health studies. The use of Semantic Web technologies such as biomedical OWL ontologies in epidemiological and public health research serve to potentially enhance the interoperability, extendibility and discoverability of research data. Inclusion of the Semantic Web technologies heading in the model also serves to encourage stakeholders into considering how well maintained, for example, links are – broken links could potentially be a sign of poor quality metadata. Additionally, by using mechanisms such eXtensible Markup Language (XML), there is the potential to create newer versions of the metadata whilst referencing previous versions; an example of where this is possible is using the DDI Lifecycle metadata standard.

In the 'Usability' model, standards can refer to metadata standards such as MIBBI or exchange standards such as HL7 or CDISC. Cross walks could either refer to mappings between metadata standards such as DDI 2 to DDI 3; and/or between clinical terminologies such as Medical Subject Headings and International Classification of Diseases. In having cross walks between metadata standards, the metadata's interoperability could potentially be enhanced. In providing cross walks between clinical terminologies, researchers wanting to run queries using multiple terminologies from disparate datasets could potentially identify which clinical codes are equivalent in support of for example, clinical cohort phenotyping. The use of repositories acts as a mechanism to enhance discoverability and provide meta-metadata. Meta-metadata are metadata about the metadata; for example, date when metadata were created, the version number and by whom.

In the 'Management and curation' model, inclusion of the 'dates' and 'versions' headings encourage stakeholders to look at the meta-metadata and encourages them to assess how timely the metadata are. Provision of this meta-metadata also helps to support version control, and scope for extension.

4.4.3.2 Novel quality assessment framework

The quality assessment framework for epidemiological and public health metadata has four components, Table 4-13:

Table 4-13 Framework component definitions

Part	Definition
1. General information	Assesses provision of, but not limited to, the types, formats and granularity of the metadata available
2. Tools and technologies	To assess the structure, application of clinical terminologies, indexing in catalogues, and use of SWTs
3. Usability	Assess presence in repositories, application of metadata standards, and provision of cross-walks or other semantic mappings
4. Management and curation	This refers to how the metadata were created and provision of and access to other metadata versions

The boxes in pink denote the area of the framework and are based on the sections of the survey. The boxes in yellow are the associated quality dimensions and the boxes in green are the question topics. This model also aims to show how the different elements are linked together and that question topics can have a basis in more than one quality dimension. In Figure 4-5, the model is broken down according to the different areas of the framework. Following development of these models and definitions, a metadata quality framework is proposed in Table 4-14.

Figure 4-5 Models of metadata quality by assessment section

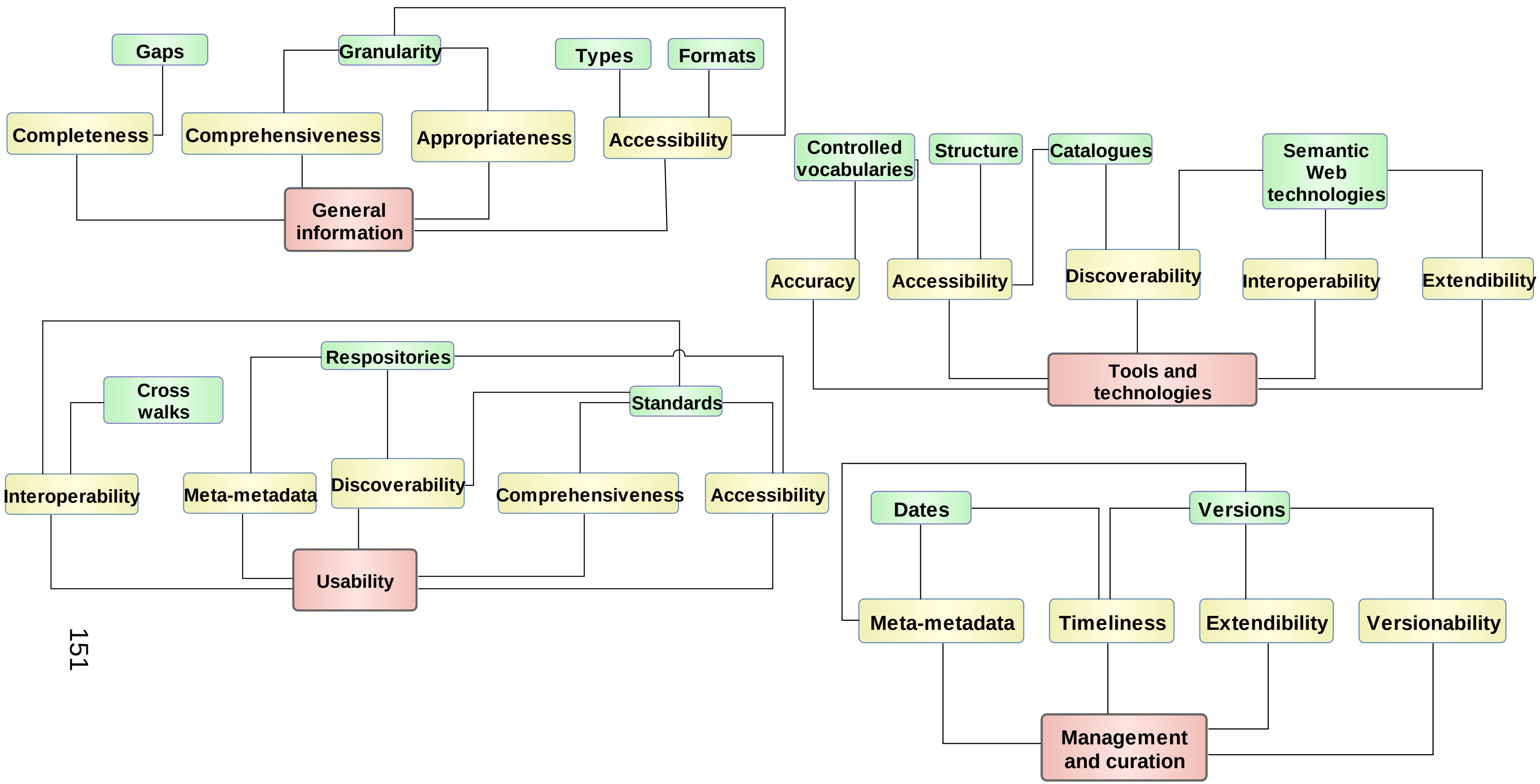


Table 4-14 Metadata quality framework

Area of metadata quality	Underlying quality dimensions	Justification	Headings
General information	Accessibility	To determine how accessible metadata are	Types of metadata
			Formats of metadata
	Accessibility, comprehensiveness, appropriateness	To facilitate comparison and ensure that metadata are available to meet a range of potential needs. Acknowledgement is given to the inability to provide an exhaustive list of different metadata levels. The levels reflect those identified as part of the review and survey.	Granularity of metadata
	Completeness	To identify where there are gaps in the metadata	Missing or incomplete metadata
Tools and technologies	Accessibility	Addresses how easily human readable versions of the metadata may be read and understood	Structure of metadata E.g. continuous prose, sectioned,
	Accuracy, accessibility	To determine extent of application of controlled terminologies.	Presence of clinical terminologies
	Discoverability, accessibility	To determine extent of catalogue use to address issues such as usability and access	Indexing in catalogues
			Restrictions on access to metadata

	Interoperability, extendibility, discoverability	To determine how extensively Semantic Web technologies have been knowingly applied and how.	Application of Semantic Web technologies
			Method of application and reason(s) for use
Usability	Accessibility, interoperability, discoverability, comprehensiveness, meta-metadata	To determine how usable the metadata are and help identify areas of potential improvement	Metadata repositories
			Metadata standards
			Cross-walks or other mappings
Management and curation	Versionability, Meta-metadata, Timeliness, Extendibility	To help improve management and curation of the metadata	Creation of metadata
			Provision of other versions

N.B. the presence or absence of timeliness information about the data fall under the ‘Management and curation’ section and in particular the two suggested headings, ‘Creation of metadata’ and ‘Provision of other versions’. Here, the user is able to record when the metadata were created – hence how current the metadata are - and if other, older versions are available.

Given the breadth of public health and epidemiology research, application of this framework would promote a more in-depth and focused assessment of the metadata. Future steps will include evaluating the framework by applying it to a series of case studies. Recommendations for changes in research data management policy and practice will be made following the evaluation.

4.4.4 Framework evaluation - Part A: Test cases

General information: In the first test case, Millennium Cohort Study (MCS), links to the UK Data service were available as were links to the variable information in addition to other such information in the Nester publisher catalogue. Metadata were available at a number of different levels and the artefacts reviewed were all complete. By applying the framework to a real world test case, I identified several areas of potential improvement. These improvements could help users to perform a more in-depth review of metadata quality. For example, in the MCS there was opportunity to browse variables and tabulate these using the Nesstar catalogue. This is an additional way to characterize the research data and mechanisms to record this kind information are needed in the framework. Also, by adding a section where additional information may be recorded, this will enable the reviewer to add findings which may not fall under a heading previously proposed in the framework. Furthermore, a section where links to other studies, sweeps and publications may be recorded is also required. For example, the MCS provided multiple links to this information but there was nowhere in the framework to record this.

Additionally, the framework did not include a section to specify which artefacts were reviewed and where these can be found. To work around this, this information was placed under the ‘Missing or incomplete metadata’ section where the completeness of the metadata were reviewed. Though not ideal, this served to record the artefacts reviewed and specify simultaneously if there were any problems. In a more low level assessment, mechanisms to specify clearly which documents are being reviewed are needed. This does however, raise issues regarding the point at which someone reviewing the metadata decides when they have gathered enough documents to review. A group of documents for each case study were randomly selected and it quickly became apparent that metadata are spread across documents and that it is through their collective review that quality may be effectively assessed. Reviewing one single document may not necessarily provide someone with enough detail to conduct the review properly. It is possible users may review multiple documents to gain a better understanding of the data. Hence, to adapt to this problem an area was be added to the general information section to record the names and locations of the research artefacts reviewed and where these can be found. By adding this section, recording this kind of information is potentially easier and clearer.

Other changes made included adding sections for variable descriptions, online data visualisation tools, links to other studies, sweeps or publications, and other. By adding these headings, I was able to record in greater detail the outcomes of the quality assessment and list where these metadata can be found. For example, in the MIDUS study I was able to record links to publications at sweep and study level. In the DNBC study, I was able to record links to all four sweeps, NCI indexed publications and a list of theses using DNBC data. No additional information was recorded under the ‘other’ heading.

When assessing the metadata for the DNBC study, locating the metadata was more challenging than MIDUS. At times there was an initial language barrier (although pages were easily translated) and links between the research artefacts were not as well established in comparison to the MIDUS study. This could be because the metadata for MIDUS were mostly found in the ICSPR catalogue which produces metadata compliant with multiple standards; while the DNBC study metadata were more sparsely located and the only metadata standard identified was DDI. The changes made enabled me to record more detailed findings from the assessment.

Tools and technologies: The structure of the metadata in the MCS was a combination of: a) PDFs comprising of continuous prose broken down by numbered paragraphs; and b) a list consisting of headings and sub-sections in the UK Data Service record, and a series of collapsible headings in the Nesstar catalogue. In the MCS, application of clinical terminologies was not found in the artefacts reviewed. This could be due to the information sitting in another document inadvertently missed; however, listings of categories and codes were located. The metadata reviewed were indexed in the Nesstar catalogue and the UK Data Service and were unrestricted. Regarding use of Semantic Web technologies, and reasons for use, DDI 2.5 compliant XML was found for the MCS through use of the UK Data Service catalogue. I could not find a justification from the organisation for using DDI compliant XML; although, given that the UK Data Service are mainly responsible for social sciences data, and DDI is a metadata standard aimed primarily at social sciences metadata, this could be a contributing factor in the decision to apply this standard.

Regarding the further development of the tools and technologies section, mechanisms to record use of category and coding schemes were also needed. Codes and categories were utilized by the test case and it is useful to record where this information may be found if there are any access issues. Furthermore, what quickly became apparent through the evaluation was the inherent links between the different sections of the review and the suggested headings, and how best to record overlapping information. For example, clinical terminologies are only one way to categorize information and can involve use of ontologies. Recording this information using the framework involves use of, the ‘Presence of clinical terminologies’, ‘Application of Semantic Web technologies’, and ‘Method of application and reason(s) for use’ headings. A potential solution here would be to treat the headings as guidelines rather than a ‘question and answer’ exercise to enable a thorough review of the metadata which can then be recorded.

For the MIDUS and DNBC studies, two changes were made following the first application: a) addition of a separate codes and categories section; and b) merging of the indexing of catalogues and metadata repositories sections. By adding the codes and categories section, I was able to review and record details regarding controlled vocabularies employed by the research teams. This is useful as in the MIDUS study I was able to record variable coding conventions and in the DNBC study I recorded the locations of the codebooks for all four interviews. By merging the catalogues and repositories sections together, I was able to reduce repetition in the framework and group together this kind of information. In the MIDUS study metadata were sourced from the ICPSR catalogue.

For both the MCS study and the MIDUS and DNBC studies, I was unable to find the method of application and reason(s) for use of Semantic Web Technologies. This could be because it is not commonplace to provide this kind of information and so it was difficult to find. I decided to include this section in the framework as I wanted to see if the method of application is linked in some way to the quality of the metadata. I decided not to remove this section before circulating the framework with stakeholders as I wanted their views on this.

Usability: In the MCS, metadata are available from both the Nesstar publisher and UK Data Service repositories both of which employ the DDI standard. Cross-walks and other such mapping could not be found in the metadata reviewed.

Enhancements to the usability section included the addition of mechanisms to record web links relating to data access and use caveats which could impact use of catalogues or repositories. Having access to this kind of information could help stakeholders better navigate through the metadata. Also, having some kind of mechanism to record provision of the underlying metadata model would also be useful. Having access to this kind of information could potentially facilitate metadata exchange enhancing the interoperability of the metadata.

Another potential enhancement was to include mechanisms to record in more detail cross-walks and other such mappings. Having access to this kind of metadata could also further support researchers' and clinicians' understanding of which variables have been mapped and the process through which this occurred. I will also remove the metadata repositories heading and combine this with the metadata catalogue heading in the previous section. This is because no new information was recorded here; rather, information was repeated from the catalogue heading.

In the usability section, changes made following the first application included addition of mechanism to record metadata models, and enhanced mechanisms to record cross-walks and other such mappings. In making these changes, it was easier to record findings which could potentially indicate the interoperability of the metadata. In the MIDUS study, four standards are used (DDI, DC, MARC21 and Datacite) whereas only DDI compliant metadata was found for the DNBC study. I could not find crosswalks and metadata models for either of the studies.

Management and citation: In the MCS, locating the creation and versioning information was easier and quicker using the underlying XML syntax; this outcome could be attributed to purpose of the metadata. Whilst the catalogue record provided a detailed description of the study; I was searching for very specific elements of the meta-metadata. A link to the XML syntax is provided from the UK Data Service record so accessing this information is relatively straightforward. However, stakeholders would need to be familiar with DDI elements to identify the information they require efficiently.

Enhancements to the management and curation section included splitting this section into two: a) the creation and version information of the metadata being reviewed; and b) the creation and version information of the review itself. In being able to record more efficiently the meta-metadata, stakeholders can quickly see how current the review is, and if another is needed. The creation and maintenance of metadata is an ongoing process and being able to conduct regular reviews will help to maintain a high level of quality.

In the management and curation section, in the MIDUS study full version history was provided inclusive of a brief description of the changes. In the DNBC study, I could not find previous versions of metadata but dates of when metadata were created were provided. Other changes made to the management and curation section included the addition of mechanisms to record the date and version of the assessment and name of the person responsible for the assessment. By making these changes, I was able to improve recording of meta-metadata of the quality assessment itself and facilitate a tracking of the different versions of the assessments. The basic tracking system with versions and dates, help stakeholders monitor the quality of the metadata across assessments and note any necessary corrective action taken and the subsequent impact. A potential weakness of this section is the lack of automation when producing the version numbers. Since this information is added manually, there is the potential for human error, version numbers could be missed or repeated as two such examples.

4.4.5 Framework evaluation - Part B Stakeholder engagement

The framework was then evaluated by 10 stakeholders (3x research data managers, 1x policy adviser and 1x metadata manager, 1x software developer, 3x post-doctoral researchers, 1x PhD student) in public health and epidemiological research. The following remarks were raised and are described below.

Firstly, the framework itself does not have a column specifically for research findings; in the evaluation part A, an altered version of the table was used. The framework should be viewed as more of a reference document and the altered version inclusive of a ‘findings’ column should be used when assessing the metadata.

The suggestion was also made to include potential answers for some of the headings; these were ‘granularity, ‘types of metadata’ and the ‘formats of metadata’. In the next version of the framework possible answers will be included to help better guide stakeholders when they are assessing metadata. These answers will be sourced from results of the review and survey.

Another query raised was the definition of certain headings e.g. “…what does “continuous prose” mean…I don’t know exactly what Semantic Web technologies refers to…” Where the structures of the metadata were being assessed, the example of ‘continuous prose’ was included. I included this example to establish whether the metadata are structured in paragraphs, tables etc. to determine how clear the metadata are to read i.e. if there are no headings/signposts, then the extent to which other researchers can navigate through the metadata quickly and easily could be negatively impacted.

Regarding inclusion of the Semantic Web technologies, this is looking at, for example, whether controlled clinical terminologies had been used as part of the study. An exemplar answer is, “the ICD 10 ontology was used as part of the study and the metadata comply with the DDI-Lifecycle schema”. Completion of this section will also help determine how clearly this information is conveyed to the person conducting the assessment. Therefore, examples will be included in the framework to help stakeholders complete the assessment more easily by reducing ambiguities.

The use of clinical terminologies in the framework was also queried; the data manager was unsure which terminologies the framework was referring to. The inclusion of the clinical terminologies heading in the framework is to deduce: a) if clinical terminologies have been used and which ones e.g. ICD 10; and b) how clinical data were encoded. This will help stakeholders to track the following: a) any version changes e.g. ICD 9 to 10 and how these changes were managed; and b) any harmonisation work to help facilitate the transitional period, on-going and further work using these encoded data. Therefore, this section will need refining to make the intention for inclusion clearer to the person assessing the metadata, and to better guide them.

Furthermore, a comment was made that finding metadata at a range of levels complete with links to publications is unrealistic and increased support for stakeholders is needed when assessing metadata. Whilst I acknowledge that the framework may provide a somewhat idealistic expectation of metadata, the framework was designed to provide a non-exhaustive list of headings to guide stakeholders. Context needs to be taken into consideration when assessing metadata quality as the granularity of metadata available may differ across the epidemiological and public health research domains.

From the evaluation, metadata standards such as DDI are commonly applied to metadata already indexed in repositories/catalogues. Though this is beneficial as the metadata elements are standardised, and there is scope to download the DDI compliant XML, this does emphasise the need for increased application of metadata standards across biomedical research and regardless of whether or not the metadata are indexed. Though this in itself presents challenges (access to the necessary resources, ongoing maintenance etc.), if the metadata are standardised, there is potential for automatic metadata harvesting protocols to be enacted and the metadata be available for automatic inclusion in a repository. This will help to enhance research data discoverability in support of improved opportunities for data reuse and repurposing.

Following the evaluation, Table 4-15 shows the finalised framework v1.0.

Table 4-15 Evaluated metadata quality assessment framework

Area of metadata quality	Underlying quality dimensions	Justification	Headings
General information	N/A	To record which artefacts were reviewed and where these can be found	Names and locations of research artefacts reviewed
	Accessibility	To determine how accessible metadata	Types of metadata *(administrative, descriptive, microdata, semantic, other)*
			Formats of metadata *(PDF, Spreadsheet, Word processed document, XML, RDF, HTML, other)*
	Accessibility, comprehensiveness, appropriateness	To facilitate comparison and ensure that metadata are available to meet a range of potential needs. Acknowledgement is given to the inability to provide an exhaustive list of different metadata levels. The levels reflect those identified as part of the review and survey.	Granularity of metadata *(research study level, single dataset/sweep of data, variable level, each time a change is made to the data, other)*
	Completeness	To identify where there are gaps in the metadata	Missing or incomplete metadata
	Discoverability	To identify how extensively metadata can be used to characterize the research data	Online data visualisation
			Provision of variable descriptions
	Comprehensiveness	To identify how well linked the metadata are to other potentially useful resources	Links to other studies, sweeps or publications
	Comprehensiveness	To record any additional details which may not fit	Other

		under any of the suggested headings. (list of proposed headings in non-exhaustive)	
Tools and technologies	Accessibility	Addresses how easily human readable versions of the metadata may be read and understood	Structure of metadata *(headings, paragraphs, tables, other)*
	Accuracy, accessibility	To determine extent of application of controlled terminologies.	Encoding and exchange standards *(Encoding - ICD, SNOMED CT, DSM, LOINC, OPCS, Read Codes, other)* *(Exchange – HL7, CDISC, other)*
			Presence of code(s) and category(ies) lists
	Discoverability, accessibility, comprehensiveness, meta-metadata	To determine extent of catalogue use to address issues such as usability and access	Indexing in catalogues/repositories *(iSHARE2, IPUMS, CALIBER, CESSDA, MRC Gateway, other)*
			Restrictions on access to metadata
	Interoperability, extendibility, discoverability	To determine how extensively Semantic Web technologies have been knowingly applied and how.	Semantic Web technologies *(biomedical ontologies e.g. ICD 10 ontology and where these can be found e.g. The OBO Foundry, XML, RDF/SPARQL, other)*
			Method of application and reason(s) for use

Usability	Accessibility, interoperability, comprehensiveness, meta-metadata	To determine how usable the metadata are and help identify areas of potential improvement	Metadata standards *(Dublin Core (or derived standard), Data Documentation Initiative Code book/Lifecycle, ISO/IEC 11179, MIBBI, Observ-OM, OME-XML, Protocol Data Element Definitions, SDMX / SDMX-HD)*
	Interoperability, meta-metadata	To determine how current the cross-walks and other mapping are and if any problems were experienced	Cross-walks inclusive of method and when these were created *(between metadata standards and/or clinical terminologies)*
			Other mappings
	Interoperability	To help improve opportunity for metadata exchange and better researchers' and clinicians understanding of the structure of the metadata	Provision of metadata model (metamodels)
Management and curation	Versionability, Meta-metadata, Timeliness, Extendibility	To manage the results of assessment and help track improvements made to the metadata	Date and version of assessment
			Name of person assessing the metadata
		To help improve management and curation of the metadata	Creation of metadata
			Provision of other versions

By structuring the framework according to the issues the suggested headings address (resulting in four sections), and not the underlying quality dimension, the dimensions repeat throughout the framework. The repeating quality dimensions demonstrate the possibility that one quality may underpin multiple suggested headings. Therefore, users should engage with the framework by working through the sections systematically using the suggested headings to guide their assessment.

4.5 Discussion

4.5.1 Literature review

A total of 11 manuscripts were identified and reviewed. As part of the analysis, the method of quality assessment and quality dimensions were recorded. Following this review, no method of metadata quality assessment was identified for use in public health and epidemiological research. Therefore, an approach to quality assessment in public health and epidemiology is needed to address this gap in knowledge. Following the review, I was able to compare the identified dimensions of quality across all the methods of assessment (Figure 4-2).

Following the identification of metadata standards in the review, the survey addressed the application of standards. The standards listed will take into account the outcomes of the discoverability study(Castillo, Gregory et al. 2014) and include other metadata standards relevant to biomedical research not previously included in the discoverability survey. Additional areas of interest which will be incorporated into the survey are the role of metadata across the research data lifecycle and the availability of software and tools and technologies to support the management of metadata. This will help inform any subsequent developmental work.

In using both biomedical and computer science databases, I identified a wide range of publications for potential inclusion in the review. This was advantageous as in areas such as e-libraries and archive management, a lot of work had been done on metadata quality assessment and so I was able to identify literature, though not included in the review, provided useful contextual information.

However, as all methods identified were from research domains other than public health and epidemiology, the applicability of the methods of assessment identified to public health and epidemiological research is inherently limited. Nonetheless, the review did identify a range of quality dimensions, such as accuracy and accessibility, which are applicable in the public health and epidemiology research domains. The challenge will be defining these taking into consideration the intricacies of public health and epidemiological research especially at a low level.

This systematic literature review identified a lack of structured methods to assess metadata quality within the context of public health and epidemiological research settings. The next step was to conduct a stakeholder survey to better understand the current role of metadata in public health and epidemiological research and which are the perceived challenges associated with the management of metadata.

4.5.2 Online stakeholder survey

Metadata types and formats: Most of the respondents who submitted data indicated that they used descriptive metadata. This outcome could potentially be a reflection of the respondents' roles in public health and epidemiology research and by extension their daily routines. Given that most of the respondents were data users employed by a university, and that metadata often accompanies public health research datasets, this could be why descriptive metadata was the most popular whilst semantic was the least popular. Further to these findings, the framework was developed in such a way that it could potentially be applied by a range of stakeholders including data and metadata producers, data managers, archivists etc. For example, data managers and archivists could use the framework as a guide to ensuring minimum metadata quality standards have been met. This is because the framework is adequately detailed to facilitate a robust and systematic assessment of metadata quality; yet, sufficiently abstract to enable applications in multiple scenarios.

Results also show that PDF was the most commonly handled format of the metadata whilst RDF was the least most commonly handled. PDFs are easily produced and managed, particularly when compared to the marking up resources using the RDF. This ease of use combined with the speed at which such documents may be produced and disseminated could be contributing factors to the popularity of this type of metadata. Furthermore, organisations such as the Health & Social Care Information Centre provide the metadata for HES data as PDF documents. Given the proclivity of researchers to use these clinical data for research purposes, familiarity with this type of metadata is ameliorated. Nonetheless, PDFs are not machine readable so this poses a problem when trying to automate the process of ingesting these documents and automating their processing.

In terms of informing the development of the model and framework, these findings can inform the framework's structure. The framework needs to take into consideration that metadata can be presented in multiple formats, and the framework needs to enable the recording of each different instance.

Research data lifecycle: The most commonly indicated stage was 'Access, use and reuse' followed by 'Analysis'. Regarding the granularity of the metadata, the most popular level was 'Research study level' followed by 'Variable level'. It is possible that given most of the respondents were data users and data providers, this could account for this result. The use of clinical data for research purposes to better inform the development of clinical policy and practice is commonplace in public health and epidemiological research. Such endeavours demand multidisciplinary teams to work even more closely to achieve research aims. It is possible that this continued need for data users capable of investigating big biomedical data coupled with the need for increased visibility of research data could have caused these survey outcomes.

In terms of informing the framework, mechanisms to record the granularity of the metadata are required. Having metadata available at different various levels could potentially help to address issues relating to comprehensibility, and presence of different levels of metadata should be recorded as part of the quality assessment. Having this kind of information recorded is helpful as stakeholders across public health and epidemiological research could more easily identify the metadata most appropriate to their needs. Potential secondary users may request variable level metadata; whereas, a member of the public may request research study level metadata. Additionally, in having a record of which metadata are available and at which levels has the potential to support metadata quality comparison work across different instances of metadata. The aim of this work would not be to ‘name and shame’ but to inform the development of quality benchmarks by which other metadata may be compared. This type of work does however raise problems relating to defining which are the acceptable levels of quality and how are these are measured. Defining these levels is made difficult by the subjectivity of such work; automatic methods of quality assessment will help to reduce any potential bias in the results of the quality assessment but again the same issue of defining what is good quality metadata arises. Therefore, the quality assessment framework will present scope for the addition of quantifiable metrics of quality, but not necessarily define these.

Creating and/or using metadata in biomedical research: Results showed that the barriers can be categorised into six categories. This outcome could potentially be explained through a previous lack of focus on the importance of biomedical metadata possibility contributing to the subsequent lack in training and support. Knowledge of the role of metadata in the research data lifecycle is increasing as is familiarity with metadata standards. Funding agencies are increasingly calling for high quality metadata to accompany research datasets and, in certain circumstances, request metadata be available in standardised, public facing catalogues. Therefore, the framework needs to have some kind of mechanism which records the overall completeness as a preliminary measure of quality; if the metadata are incomplete, the first task stakeholders must address is to record the missing metadata before the rest of the metrics may be applied.

Tools and technologies: In identifying how respondents chose a tool and/or technology provided a clearer understanding of which factors influence decisions made. Results show the most commonly indicated way of selecting a tool and/or technology was standard practice. This result is most likely due to stakeholders’ tendency to refer to best practice guidelines to inform the undertaking of research subsequently causing recommendations to gradually become standard practice. This is important to the development but more so the implementation of the model and framework as integrating quality assessments of metadata in best practice guidelines and assessments becoming standard practice is a future goal of this work.

Clinical terminologies and classification systems: Of those who provided data, the terminology most respondents had come across was ICD followed by MeSH. ICD was developed to be used as part of clinical coding for death certificates whilst clinical terminologies such as Read Codes and SNOMED CT were designed to encode information for clinical care. Both sources of data are routinely harnessed in public health and epidemiology for secondary use but the current proclivity to encode clinical information using ICD, could be a contributing factor to the results of the survey. Familiarity with MeSH could possibly be attributed to stakeholders' use of databases such as PubMed which use MeSH terms for indexing purposes.

Encoding standards, which standardise free text found in clinical documents using controlled clinical terminologies to systematically organise information and support knowledge management are used in clinical practice to primarily encode clinical information; and yet their use in clinical research is often to facilitate querying and clinical phenotyping using ontologies. Using SWTs in clinical research can potentially increase interoperability and extendibility of metadata in addition to enhancing research opportunities for the research data. In terms of informing development of the framework, scope must be included to record presence of encoding standards, how these were implemented and for which purpose(s).

The participants were then asked to share any difficulties experienced with these clinical terminologies. Challenges included a lack of medical knowledge and the lack of ease of use. Given that medical knowledge is needed to better understand the encoding process, when querying for data using clinical codes, this can become difficult if there is a lack of medical knowledge. Often researchers work in multidisciplinary teams which include data users and clinicians to determine which codes are needed. And given that mostly data users responded, this could have caused this outcome. Furthermore, use of clinical terminologies is not mandatory in clinical research as it is in clinical practice, therefore, availability of clinical terminologies will inevitably vary. This could also have been a contributing factor to the results of these survey questions.

With regards to informing the framework, there needs to be mechanisms to firstly record any use of clinical terminologies; secondly – any issues such as lack of definitions which could impact subsequent use; and thirdly, a note describing which version of a clinical terminology was used. This is to help bring any (potential) secondary users to a greater understanding of how any clinical information was encoded.

Catalogues: The issues associated with the use of metadata catalogues could possibly be attributed to the lack of academic incentives to use this kind of catalogue coupled with limited publicity of the existence of such platforms. Metadata catalogues serve as mechanism to enhance the discoverability of data and given recent motivation to adopt a cyclical approach to the use of data, being able to identify and characterise data without accessing the research data directly is potentially beneficial. However, catalogues are not extensively used in this way and in certain circumstances researchers continue to contact data managers to ascertain variable level information.

In terms of informing development of the framework (Table 4-14), mechanisms to enhance discoverability need to be identified, and where possible, record the details of any metadata catalogues used for indexing.

SWTs: These findings could be a reflection of the lack of awareness of the use of SWTs in public health and epidemiology research and the need for a greater understanding of the supporting technologies which help enable public health and epidemiology research.

Metamodels: With regards to use of metamodels, 5 people said they used some kind of metamodel examples of which include HL7 and OpenEHR. Again, limited publication of metamodels could be attributed to a lack of awareness of mechanisms such as these and the critical role they play in enabling clinical research and practice.

In terms of informing development of the framework, there needs to be a mechanism to record use of these technologies and the purpose. For example HL7 is an exchange standard used primarily to facilitate the sending of encoded messages in secondary care. Recording use of HL7 or the underlying reference information model could enhance provision of contextual information and better support any subsequent analysis work. Recording the use of XML, to provide a machine-readable format of the metadata, could help to support stakeholders in identifying how interoperable the metadata are and identify any areas of potential improvement.

Metadata usability: Given that metadata needs to be accessible to potential secondary users to enable them to characterise the research data, these results serve to emphasise the importance of the provision of properly indexed, usable metadata. For example, only listing variables is not entirely useful, information round how these variables were collected, when, why etc. is needed to provide the context around these variables and to enable these metadata to be useful; without these additional metadata the variable metadata is not as usable as it could be.

However, there remains a tension between having enough metadata to characterise the dataset and the amount of resource required to generate the metadata (Ellul et al., 2013). Of those who submitted additional aspects of importance to metadata usability, suggestions included, but not limited to, multiple formats of the metadata and semantic mappings. As metadata standards such as DDI are being continuously developed, semantic mappings between the different schemas are needed to support interoperability.

Therefore, the framework needs to enable the person assessing metadata quality to record use of metadata repositories, metadata standards, provision of human and machine readable versions of the metadata, and use of any crosswalks. Recording information such as this will help evaluate quality aspects such as accessibility, interoperability, meta-metadata and discoverability.

Metadata quality assessment: A challenging aspect of quality assessment is being able to define the dimensions of quality at a low level. The definitions presented in the survey are generic and potentially applicable across all aspects of public health and epidemiology research. However, the difficulty remains in being able to deliver pragmatic dimensions for the sub-domains of research. For example, accuracy of metadata associated with EHRs may not necessarily be the same as for metadata associated with cohort studies; nor is the approach to quantifying accuracy and the mechanism through which it may be tested. In trying to provide low level definitions of the quality dimensions, context must be taken into consideration to help improve implementation of the quality framework. Therefore, the initial framework should serve a preliminary evaluation of quality and the catalyst for more in-depth analysis of the metadata handled. In terms of informing development of the framework and its subsequent evaluation, low level definitions will be suggested for case studies to evaluate how well the framework lends itself to expansion within different contexts.

Based on the outcomes of the review, the framework needs to specify which aspects of quality are being assessed and have a way of recording the outcomes, and if necessary, steps to be taken if problems with the metadata are identified. This will also help to encourage the recording of meta-metadata and improved management of the different versions of the metadata. Furthermore, when developing the framework for public health and epidemiology, some kind of mechanism enabling researchers to decompose the assessment into a series of smaller assessments is needed. The subsections could potentially be based on the different areas of metadata quality identified and used to structure the survey. In having a standardised approach to metadata quality assessment, this could potentially help improve stakeholders’ ability to compare across metadata instances through the standardised recording of the assessment outcomes.

4.5.3 Framework

The framework is based on all 10 identified quality dimensions in biomedical research but focus was placed on basing the questions on the top 5 scoring quality dimensions as voted for in the survey. The top 5 were: accuracy, accessibility, discoverability, appropriateness and comprehensiveness. The following describes how these may be assessed:

Accuracy: stakeholders could establish how timely the metadata are and when the metadata were updated. It is also possible to look at any application of clinical terminologies, in particular looking at how these were applied and if any code listings have been provided.

Stakeholders could determine if sufficient metadata have been provided or whether more is needed to better understanding of the research data and its context. Nevertheless, determining the level of sufficiency is challenging as this could vary depending on a person's point of view – what is sufficient to one, may not be sufficient to another. In this case, the framework itself would not judge sufficiency, this would be determined by the user.

Accessibility and discoverability: stakeholders could determine how well structured the metadata are and if they are available through a catalogue. Stakeholders are advised to establish whether encoding standards such as ICD, SNOMED CT etc. and/or exchange standards such as HL7, CDISC etc. have been utilised. It is also possible to assess how well knowledge has been managed through effective application of biomedical ontologies and whether these can be viewed possibly through an online portal such as The OBO Foundry, The Open Biological and Biomedical Ontologies.

Appropriateness: stakeholders could determine whether necessary details have been provided to enable other stakeholders to understand and potentially reuse the research data. These details (under appropriateness) differ from sufficient (under accuracy) as the details provided could be accurate but inappropriate in the sense that they do not provide the information needed by the user – the metadata is accurate, but does not answer the user's question.

More specially, stakeholders are advised to determine how comprehensible the metadata are at different levels; for example, stakeholders could assess whether sufficient descriptions have been provided of variables to enable potential secondary researchers to characterize the dataset.

Comprehensiveness: this could be determined by establishing how encompassing the metadata are and if the metadata contain gaps. For example, stakeholders could assess whether details of changes made to a sweep of data have been communicated effectively.

However, the difficulty remains in being able to deliver pragmatic dimensions for the sub-domains of epidemiological and public health research. For example, accuracy of metadata associated with electronic health records may not necessarily be defined in the same way as for metadata associated with cohort studies; nor is the approach to quantifying accuracy and the mechanism through which it may be tested. In trying to provide low level definitions of the quality dimensions, context must be taken into consideration to help improve implementation of the quality framework.

4.5.4 Framework evaluation

Metadata standards such as DDI are commonly applied to metadata indexed in repositories/catalogues. Though this is beneficial as the metadata elements are standardised, and there is scope to download the DDI compliant XML, this does emphasize the need for increased application of metadata standards across epidemiological and public health research regardless of whether metadata are indexed. However, this in itself presents challenges such as having access to the necessary resources and assigning responsibility for ongoing maintenance.

Furthermore, for each test case, I was unable to find the method of application and reason(s) for use of Semantic Web Technologies; potentially as it does not seem commonplace to provide this kind of information. I had decided to include this section in the framework as I wanted to see if the method of application is linked in some way to the quality of the metadata. As stakeholders are able to customize the framework to suit local needs, I decided to keep this section as it is not a mandatory field in the assessment.

A potential weakness of the framework is that given multiple research artefacts may need to be sourced before quality can be assessed, the question of how much metadata is needed is raised. Another potential weakness of the framework is the lack of quantitative analysis; all assessments are manual/qualitative which are inherently subjective.

4.6 Conclusions

The use of epidemiological and public health data for research purposes to better inform the development of clinical policy and practice is critical in public health and epidemiological research. One of the main challenges in assessing quality in epidemiological and public health research is a lack of awareness of the issue of poor quality metadata and the potential implications this can have on research data discoverability. Improved awareness of the issue of metadata quality is needed, as are mechanisms to integrate metadata quality assessments into daily routines of stakeholders in epidemiological and public health research.

A novel framework was created and evaluated as a platform-independent method of assessing metadata quality, with the goal of improving metadata in epidemiological and public health research settings and enhancing the potential for data discovery and reuse in the context of epidemiological and public health research studies.

My next steps include engaging with stakeholders to establish a set of requirements for a series of computational metrics. The short term goal is to identify a set of quantitative measures of quality to compliment the framework. The longer term goal is to use these metrics to increase objectivity and automate/quicken the overall assessment process.

4.7 Summary of major findings

4.7.1 Literature review

A total of 11 publications were eligible for full review; none of which are aimed specifically for use in epidemiological and public health research settings. A total of nine different quality dimensions were also identified. The literature review highlighted the need to create a framework for epidemiological and public health metadata.

4.7.2 Online stakeholder survey

Results showed that the most common type was administrative (31%) and the most popular format was PDF (23%). The least popular format was RDF with only 5%. Survey results show that most respondents used metadata during the ‘analysis’ stage of the RDL. The most popular level of metadata as indicated by the respondents was ‘research study level’ with 28%.

I identified the main barriers to creating and/or using metadata in biomedical research include: a) lack of skills/experience in generating/using metadata; b) inconsistencies (namely formatting metadata); c) inadequate tool availability; d) standards; e) ethics; and f) inadequate resources (time and cost).

Results also showed the most common response indicated suggestions from colleagues determined how they selected tools and technologies. Survey results also show that ICD was the most popular statistical classification system followed by MeSH. Challenges associated with use of clinical terminologies can be categorised into: a) lack of medical knowledge impacting meaningful use; b) ease of use; and c) inconsistent availability. The results showed that ‘Accuracy’ was the most important quality of biomedical metadata. The least important was ‘meta-metadata’. Results show that 19 respondents sometimes assess the quality of metadata whilst eight respondents never do so.

Furthermore, of the 28 respondents who submitted data, only one respondent indicated they used a quality assessment criterion. Moreover, challenges associated with assessing metadata quality in biomedical research include: a) lack of guidance and awareness of the metadata quality issue; b) lack of domain-specific knowledge negatively impacting on how well the metadata are understood; c) limited tools availability (supportive software and being able to access these); and d) limited resources (time).

4.7.3 Model/framework and evaluation

These models are: a) general information; b) tools and technologies; c) usability; and d) management and curation. These models underpin the novel metadata quality assessment framework for use in epidemiological and public health research settings. The framework was validated and evaluated through iterative application to the metadata for a series of test cases and areas for improvement were identified; corrective action was taken to address these.

4.8 Chapter summary

This chapter focused on reviewing existing methods of metadata quality assessment, investigating the current state of the art in epidemiological and public health assessment role of metadata quality in public health and epidemiology research, and creating and evaluating a framework to assess metadata in public health and epidemiological research settings.

This study involved a literature review which identified of 11 studies and nine dimensions of quality; none of which were aimed specifically at biomedical metadata. The online survey identified use of metadata across the research data lifecycle but quality assessment was conducted by most people only sometimes. The online survey also confirmed the lack of metadata quality assessment frameworks for use in public health and epidemiology and confirmed the need address the gap in metadata management.

Based on the results of the literature review and online survey, a model of epidemiological and public health metadata was created. These were then validated and evaluated using a series of test cases and engaging with stakeholders. The framework can assist stakeholders in assessing epidemiological and public health metadata quality in a systematic and robust manner.

Chapter 5 Research case study 3: Improving the recording of consent for record linkage metadata in longitudinal studies

5.1 Introduction

In the previous two chapters I presented the data discoverability and metadata quality studies. Findings from both these studies have shown that having access to metadata is vital in enabling stakeholders to undertake record linkage research. Researchers are increasingly linking combinations of longitudinal cohort studies, genomic, and administrative datasets together to produce enriched datasets through their inherent complexity.

The longevity of these data is such that it must be matched with consent models equally as enduring. However, developing these models is a challenging process hindered by a lack in standardised methods of recording them. There is guidance on the format and wording of consent but this guidance changes frequently and there is a lack of standardised methods to record this process(Friedlander, Loeben et al. 2011; National Research Ethics Service 2011). According to a study by Rothwell, Wong et al. (2014) there are currently no standardised formats to present the different elements of consent and there are additional problems such as a lack of recorded comprehension associated with the consenting process.

Case study: Life Study

Life Study was a birth cohort study based in the UK and aimed to follow 80000 babies through to adulthood[1].

The Life Study website provides a list of downloadable resources for stakeholders to explore and use. The difficulty in using these however lies in their format – the resources are in PDF and there is currently no standardised method of recording these. Consequently, resources such as the consent form to access personal records remains in an unstandardised format.

The aim of this research case study is to address this current lack of standardised methods of recording consent for record linkage by creating and evaluating a novel metadata management model. This model can be applied to consent forms from studies such as Life Study to record the elements of consent in a standardised way.

1. http://www.lifestudy.ac.uk/

In this chapter I present the recording consent for record linkage study. In this study I created and evaluated a novel metadata management model to support improved recording of consent for record linkage in consented longitudinal studies. Please see 1.6.3 for a detailed description of my role, responsibilities and contributions to this study and the published report.

5.2 Informed consent in epidemiological and public health research studies

Informed consent is a fundamental aspect of the clinical research process and to enabling record sharing and linkage. Consent is required for data from different sources to be linked by researchers for epidemiological studies. When obtaining consent, an individual can consent or on behalf of another; for example, a legal guardian/parent consenting on behalf of an infant in a longitudinal study. In such studies, the extent to which a legal guardian/parent consents on behalf of a child lessens as the child matures and their autonomy increases (the extent of assent and dissent taken into consideration) (Hens, Van El et al. 2013). Researchers based in the UK, can request permission from the National Information Governance Board to suspend the requirement of consent(Knies, Burton et al. 2012). Section 251 allows researchers to access data without consent under certain circumstances; for example, it being impractical to gain consent from potential research participants (NIGB 2013).

The consent process can be modular in design enabling participants to partake in a study but opt-out of certain tests and/or analyses such as whole genome analysis (Buckow, Quade et al. 2014). Consent may be withdrawn at any point during a study and can occur on different levels such as: a) complete removal of consent – all data must be destroyed and its use discontinued; b) removal of consent for future contact - allows continued use of pre-existing data and record linkage; and c) removal of consent for future contact and record linkage – allows continued use of pre-existing data only (Ries, LeGrandeur et al. 2010).

Consent for record sharing and record linkage is a multileveled process and all decisions made by potential participants must be well informed if stakeholders in research are to be entrusted with the participants' records. Though use of linked health records enables researchers to yield a greater insight into life course influences, concerns over the security of these data and with whom the data is being shared, are raised. If potential participants fear their data will not be handled with diligence and respect, whilst acknowledging there is no such thing as guaranteed complete security, they may be discouraged from sharing their records for research use.

5.2.1 Challenges associated with recording consent for record linkage

When collecting research data as part of epidemiological and public health studies, researchers each have a unique way of managing the consent process. When linking different sources of data together, these different consent models must be harmonised before researchers can use these linked datasets. Harmonising the different approaches to consent, in addition to methods for archiving and preserving data, can help to stream line research processes at a national level (Singleton and Wadsworth 2006). By simplifying and standardising these approaches, there is the potential for researchers to better understand the extent to which consent has been given by participants and under which conditions.

The provision of metadata can assist consent model harmonisation and the need for enhanced application of extended metadata standards is increasing (Pisani and AbouZahr 2010). Access to standardised, and where possible, automatically generated metadata can support stakeholders in characterising datasets and identifying the infrastructure needed to enable effective use of the research data. Examples of consent details include, information around which records could potentially be linked together e.g. health and education, the format of the consent form itself e.g. the contents of the form i.e. the questions, instructions and confirmatory statements, the logic, and how the consent may be obtained e.g. through face-to-face interview or electronically through an online form. Therefore, an increased investigation into the adoption of health information standards to facilitate enhanced use of big biomedical data for research purposes is needed and builds on recommendations which actively encourage the use of standards (Boulton, Campbell et al. 2012; MRC 2014).

Currently, there is a gap in knowledge around the recording of consent for record linkage in longitudinal studies using information standards. There is guidance on the format and wording of consent but there is a lack of standardised methods to record this process (Friedlander, Loeben et al. 2011; National Research Ethics Service 2011). By having a standardised method of recording consent, stakeholders will potentially be better placed to compare across the different consent models and will also help to reduce ambiguities when trying to establish the extent to which consent has been given (Administrative Data Taskforce 2012).

5.3 Study aim and objectives

The aim of this work was to create and evaluate a novel method of recording consent for record linkage metadata applicable to longitudinal studies. This study had four objectives: a) systematically identify and review current methodologies for recording consent for record linkage using metadata elements in the context of bespoke investigator-led consented longitudinal studies; b) systematically identify and extract the key elements of consent for record linkage in longitudinal consented studies; c) critically evaluate DDI 3.2; and d) create and evaluate the metadata management model by iteratively applying it to a series of test cases.

5.4 Ethics

Ethical approval was not required.

5.5 Methods

5.5.1 Literature review

To perform the literature review in a methodical and thorough manner, I used the PRISMA checklist(Moher, Liberati et al. 2009) for guidance. This systematic literature review sought to answer the following questions:

1. What developmental work has been undertaken on models to record consent forms for record linkage in longitudinal studies?
2. What methods were used to develop the model?
3. What evidence is available regarding the successes or failures of different approaches?
4. Are there any best practices or guidelines available to assist with development?

These questions were relevant to this research case study as the results informed the direction in which I took this work. By knowing what developmental work had already been undertaken, and the methods used, I was able to establish what had already been achieved and identified potential gaps in knowledge. Further, by gathering the evidence for the success and failures of the approaches and establishing best practices or guidelines, this information informed the methods used in designing and developing the metadata management model.

5.5.1.1 Eligibility criteria

To be included in the review, the literature had to be available in English and the full text available through open access sources or using institutional login. All forms of media were accepted and the latest date of publication was limited to January 2015.

5.5.1.2 Information sources and search terms

The review was conducted in January 2015. I used the following databases, PubMed, Ovid, Scopus, The Cochrane Library, JSTOR, ACM Digital Library, Lecture Notes in Computer Science, Web of Science, Inspec, Google, Google Scholar, Intute, and forward citation tracking(Kuper, Nicholson et al. 2006). I used the following search terms: ‘consent forms’, ‘longitudinal studies', 'record linkage', 'informed consent' and 'consent models' with the Boolean logic 'and'; 'or'.

5.5.1.3 Study selection

When publications were identified, the citation information was downloaded into the reference management software EndNote. Using this software, duplicates were identified and removed.

5.5.2 Qualitative analysis of consent forms: Metadata management model design and development

I designed and developed the metadata management model in two steps. Firstly, I used 30 consent forms from nine longitudinal studies to inform the design and development of the model. (Table 5-1). The elements were grouped according to theme; these themes were collated inductively and iteratively. The four resulting themes were: a) people; b) consent form; c) personal records; and d) information document. I then combined the elements for each section and removed repeating elements to reduce redundancies in the model. I created the model using Unified Modelling Language. The Unified Modelling Language (UML) is an open modelling standard which helps users to model a domain and identify key processes and people (Brazma, Krestyaninova et al. 2006). UML is often used to develop object oriented models to support object oriented systems design and analysis. Object orientation is a technique used to decompose information, processes, attributes and behaviours into manageable subunits. These subunits are referred to as objects and each object has a unique name and can have variables (the attributes) and methods (the behaviours). When creating the metadata management models, I used the object oriented principle of inheritance. This involves a class, known as child or sub-class, obtaining variables and methods from another class, known as the parent or super-class. Inheritance of such methods and variables potentially increases efficiency as the risk of repetition is reduced.

The studies selected to inform design and development of the model had to be investigator-led, longitudinal and consented, be population focused, and have a record linkage component. The studies were identified through a combination of desk research and engaging with stakeholders. The studies were: Avon Longitudinal Study of Parents And Children (ALSPAC) (2016b), Born in Bradford (2016c), British Household Panel Surveys (2016d), Health Survey for England (Mindell, Biddulph et al. 2012), Life Study (2016l), Millennium Cohort Study (Connelly and Platt 2014), Scottish Health Surveys (Gray, Batty et al. 2010), UK Biobank (2016p) and Understanding Society (2016q).

Table 5-1 Longitudinal, consented studies

Study	Years	Coverage	Participants	Consent forms reviewed	Record linkage
ALSPAC Supplementary Figure 1	1991-on-going	Bristol and nearby areas	14,062 children, 14,541 mothers	1. Project to Enhance ALSPAC through Record Linkage	• Health • Education • Benefits and earnings • Police
Born in Bradford Supplementary Figure 2	2006-2011	Bradford	13, 857 children, 12,453 mothers	1. Father's consent form 2. Mother's consent form 3. Born in Bradford Allergy and Infection Study – Mother's consent form 4. Mechanisms of the Development of Allergy – Mother's consent form	• Health
British Household Panel Survey Supplementary Figure 3	1991-2009	Britain	10,300 individuals	1. Form B (all households with children aged 0 to 15 years) - Adding information from administrative health records 2. Form C (all adults aged 16-24) - Adding information from other sources 3. Form D (all adults) - Adding information from mother sources 4. Form E (all households with a child aged 3 to 15 years) - Adding information from other sources	• Health • Education • National Insurance contributions • Benefits and tax records • Saving and pensions
Health Survey for England Supplementary Figure 4	1991-on-going	England	4,000 children, 16,000 adults	1. NHS Central Register and Cancer Register – (Adults 16+) 2. Hospital Episode Statistics – (Adults 16+)	• Health

Study	Years	Coverage	Participants	Consent forms reviewed	Record linkage
Life Study Supplementary Figure 5	2014-2015	UK	Expected to have 80,000 babies born between 2014 and 2018	1. Consent form for partner at 28 week visit (MC) 2. Consent form for pregnant mother at 28 week (MC) 3. Consent form for record linkage at 4 months (child) 4. Consent form for child at 4 month visit (NC) 5. Consent form for child at 4 month visit (MC) 6. Consent form for partner at 4 month visit (NC) 7. Consent form for mother at 4 month visit (NC) 8. Consent form for record linkage at first visit/contact (father/partner) 9. Consent form for record linkage at first visit/contact (mother)	• Health • Education • Mobile phones (to establish how often calls are made) • Economic • Fertility (only if consent is specifically given) • Potentially records held by the Department for Work and Pensions
Millennium Cohort Study Supplementary Figure 6	2000-on-going	UK	19,519 children, 19,244 families	1. Age 7 Survey – Information from other sources	• Health • Education
Scottish Health Surveys Supplementary Figure 7	1995, 1998, 2003, 2008-2015	Scotland	Approximately 2,000 children, 6,5000 adults from 4,500 households	1. Scottish Health Records – (Adults 16+) 2. Scottish Health Records – (Children 0-15) 3. Scottish Government Follow-up Research – (Adults 16+) 4. Scottish Government Follow-up Research – (Children 0-15)	• Health
UK Biobank Supplementary Figure 8	2006-2010	UK	503,316	1. Consent Form: UK Biobank	• Health
Understanding Society Supplementary Figure 9	2009	UK	40,000 households	1. Form A: Adding information from administrative health records – adults (16+) 2. Form B: Adding information from administrative health records – children (0-15yrs)	• Health • Education • Economic circumstances

Study	Years	Coverage	Participants	Consent forms reviewed	Record linkage
				3. Form C: Adding information from administrative education records – adults (16-24) 4. Form D: Adding information from administrative education records – children (4-15yrs)	• Transport

5.5.3 DDI3.2 critical evaluation

Secondly, I critically evaluated the Data Documentation Initiative 3.2 (DDI 3.2) – the prevailing existing metadata standard - to determine its applicability to epidemiological and public health research settings. More specifically, I investigated the extent to which consent for record linkage in longitudinal study elements can be recorded using DDI 3.2.

I critically evaluated DDI3.2 by systematically mapping manually, the elements in DDI 3.2 to the consent elements previously identified. By creating these mappings, I was able to determine whether a direct link could be made between consent elements and DDI3.2 elements; or, if a related element in DDI3.2 could be used, potentially in conjunction with the 'Note' element, to record the consent element. This critical appraisal also enabled me to determine which consent elements could not be recorded using DDI 3.2 thus identifying insufficiencies in the metadata standard and where potential extensions are needed to create low level descriptions.

5.5.4 Metadata management model evaluation

I evaluated the metadata management model by iteratively applying it to the metadata from three consent forms as test cases making any necessary changes for improvement after each application. The first test case was, the English Longitudinal Study of Ageing (ELSA) (Steptoe, Breeze et al. 2013). The consent form was taken from wave six and is entitled, 'HES and DWP consent form'. The second was the Canadian Longitudinal Study of Aging, Étude longitudinale canadienne sur le vieillissement (CLSA) study. The consent form the CLSA study is the 'Consent form – Home Interview & data Collection Site Visit'. The consent form requests consent to access and link to health data. The third test case was, *Growing Up in Australia:* The Longitudinal Study of Australian Children (LSAC). The consent form selected from LSAC is the 'Adolescent study participant form' taken from wave seven. This consent form requests access to health records and pharmaceutical benefits information. In adopting this approach to model evaluation, I was able to determine how fit for purpose the model was and which are its strengths and weaknesses.

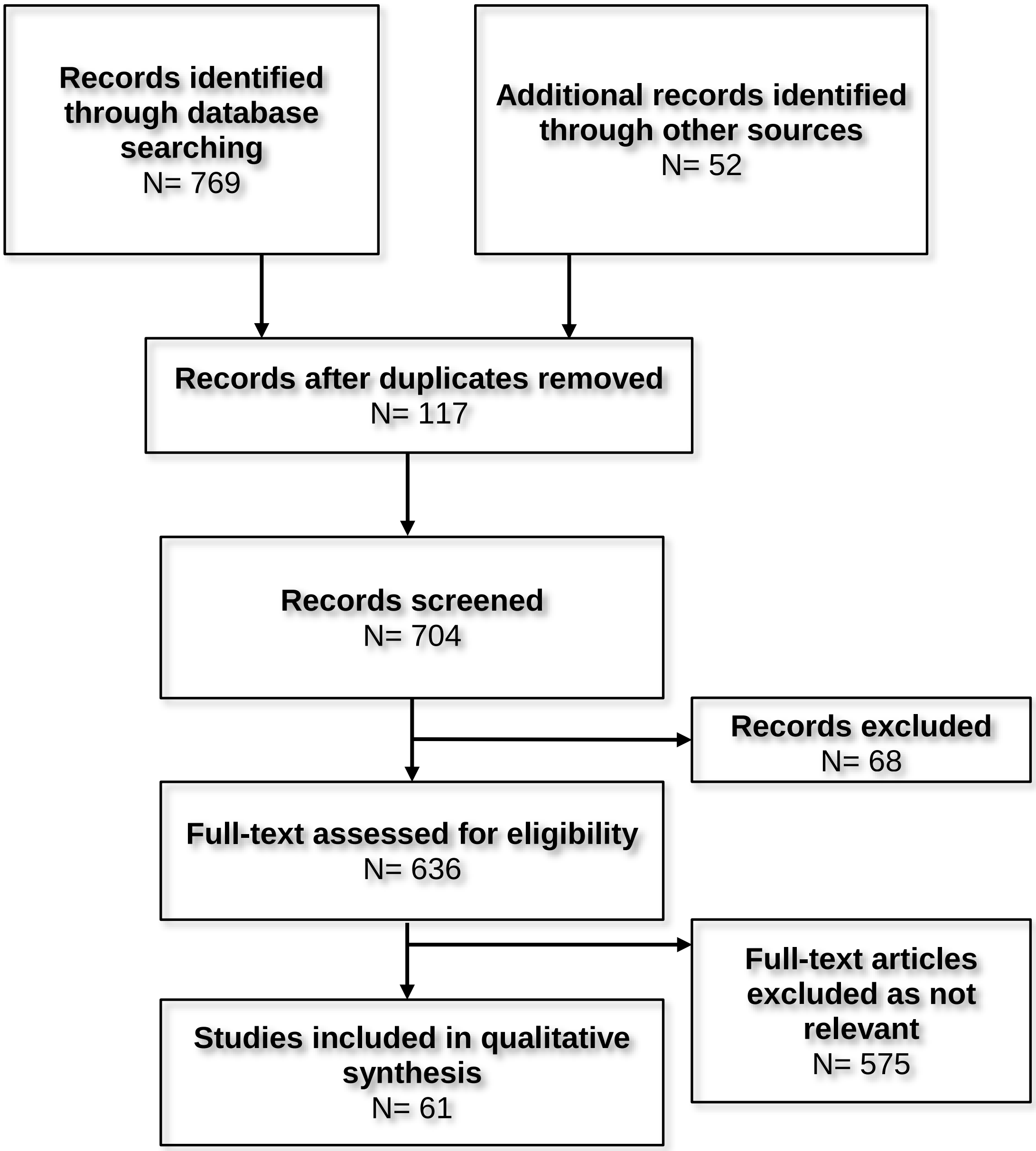
5.6 Results

5.6.1 Literature review

5.6.1.1 Study selection

A total of 61 manuscripts were identified and reviewed, Figure 5-1.

Figure 5-1 PRISMA flow diagram


Records identified through database searching
N= 769
Additional records identified through other sources
N= 52
Records after duplicates removed
N= 117
Records screened
N= 704
Records excluded
N= 68
Full-text assessed for eligibility
N= 636
Full-text articles excluded as not relevant
N= 575
Studies included in qualitative synthesis
N= 61

5.6.1.2 Synthesis of results

I synthesised the results by identifying the primary theme of the manuscript and categorised it accordingly. Seven themes were collected inductively and iteratively: a) analysis - to either improve or establish knowledge; b) comparison of models – comparison of different types of consent; c) consent aspects of secondary uses of data: discussion of consent in the research context; d) development of tools to assist consent process - development of methods to assist with management; e) discussion of a single model - discussion of a single method of requesting consent; f) establishing and/or improving participant understanding - development and/or discussion of methods/tools to better patients' understanding; g) development of a new model/form and h) other.

Table 5-2 Manuscripts fully read and categorised according to theme

Theme	Percentage
Analysis	20%
Comparison of models	25%
Consent aspects of secondary uses of data	3%
Development of a new model of consent/form	8%
Development of tools to assist consent process	3%
Discussion of a single model	7%
Establishing and/or improving participant understanding	14%
Other	20%

5.6.2 Metadata management model design and development

Having reviewed the literature, I then qualitatively analysed the consent forms, Table 5-1. Results of these analyses identified four main groups of metadata elements:

a) **people** – those involved in the consent process
b) **consent form** – the composition of the consent form
c) **personal records** – potential sources of personal data e.g. health, education etc.
d) **information document –** additional informational documents accompanying the consent forms.

The following tables present the metadata elements for each of the nine longitudinal studies Supplementary Table 15 – 22. The following sections describe the four main groups of metadata elements:

5.6.2.1 People

The first section refers to all those involved in the consent process. Following the analyses, common demographic details included: full name, dates and signatures of those consenting, statement of whether consent was on behalf of a child, confirmation of understanding and details of how to withdraw consent. The persons most commonly involved in the consent process is the person consenting (interviewee), and in certain situations, the person for whom consent is given if this is different from the interviewee. Others included the interviewer or staff member, as referred to in UK Biobank, teacher in the MCS, and GP in Life Study.

For the ALSPAC and UK Biobank studies, consent could not be given on behalf of another. Whereas, in certain consent forms for, BHPS, MCS, Scottish Health Survey, Understanding Society, Born in Bradford, and Life Study, a parent/guardian could consent on behalf of a child or children. It was also noted that the consent form completed by (potential) participants sometimes depended on age and/or relationship to the person for whom consent is given. This was reflected in the consent forms available for the BHPS, Health Survey for England, Scottish Health Survey, MCS and Understanding Society - the consent form completed depended on age. For Life Study, the consent form completed depended on relationship and stage of pregnancy/number of months post birth. A combined set of elements can be found in Table 5-3.

.

Table 5-3 Combined elements for persons identified

People	
Persons	**Demographics**
• Person giving consent o Individual consenting for themselves o Parent/guardian on behalf of a child • Details o Date o Location • Interviewer o Name ▪ Forename ▪ Surname • Witnesses o Name ▪ Forename ▪ Surname • Identifiers o NHS number o Passport number o Other	• Name o Forename o Surname • Birth details o Date o Location • Address • Ethnicity • Nationality • Disability • Contact details • Next of kin

5.6.2.2 Consent form

The second section, consent form, focuses on recording the composition of the consent form. Following the analyses, the consent forms all followed a similar basic pattern in terms of structure. Each had the consent questions/statements and then confirmation of consent through the signing, dating and printing of name on the form. Each of the consent forms followed a particular logic and specified to which organisations they were associated. Furthermore, studies such as Understanding Society and BHPS provided full contact details; whilst Life Study specified data holders e.g. the English Department for Children, Schools and Families. The consent form for UK Biobank however, does not specify from where the health records could be obtained. Other studies provided details of which data may be accessed e.g. health records (Born in Bradford/Scottish Health Survey/Health Survey for England), health and education (Understanding Society/MCS/ALSPAC) and economic, education and health (BHPS). The combined elements can be found in Table 5-4.

Table 5-4 Combined consent form elements

Consent form	
General	**Consent statements/questions**
• Description/aims • Undertakings o Declaration o Rights of participants • Organisations/ Data providers o Funding agencies o Universities o Government departments o Archive • Confirmatory information o Confirmation of understanding o Signature o Date o Full name	• Method of collection o Computer assisted personal interviews (CAPI) o Computer assisted self-interviewing (CASI) • Questions/consent statements o Logic o Purpose o Potential Reponses o Codes and categories

5.6.2.3 Personal records

The third section, personal records, refers to the records of individuals consent may be given for sharing and/or linkage purposes. I found that the most common type of record to link to was health where consent was requested across the three sectors of care; the least common type was mobile phone record. It was also noted that the only study to request police records was ALSPAC. Regarding the education records, the combined elements included all three education sectors and their associated data providers such as the Department of Education. The organisation most commonly cited as a potential provider of data was the NHS.

Studies which requested access and use of educational records included: ALSPAC, BHPS, MCS, Understanding Society and Life Study. Access and use of economic records was only requested in the Life Study, BHPS and ALSPAC studies. Request was made to access data on, for example, National Insurance contributions and participation in any benefits programmes. The combined elements can be found in Table 5-5

Table 5-5 Combined elements for personal records

Personal records				
Health	**Education**	**Economic**	**Criminal**	**Mobile**
• Organisation ○ National Health Service ▪ The NHS Information Centre ▪ NHS Central Registrar ○ Department of Health ○ General Registration Office ○ Office for National Statistics • Healthcare professional ○ Primary care ▪ GP ○ Secondary care ○ Tertiary care • Clinical Terminologies ○ ICD-10 ○ ICD for Oncology ○ SNOMED-CT ○ Read Codes ○ DSM ○ OPCS	• Type ○ School records ○ Further education ○ Higher education • Organisations/Data providers ○ Department of Education ○ The Data Service ○ Department for Business, Innovation and Skills ○ Universities and Colleges Admission Service (UCAS) ○ Higher Education Statistics Agency ○ Department for Children, Schools and families ○ Department for Children, Education, Lifelong Learning, and Skills ○ Government Education Directorate ○ Department of Education/Education and	• Organisations ○ Department for Work and Pensions ○ HM Revenue and Customs • Records ○ Salary ○ National insurance contributions ○ Tax ○ Savings ○ Benefits ○ Pensions	• Organisations ○ Ministry of Justice • Records ○ Official cautions ○ Convictions	• Past • Current • Future

Personal records				
Health	**Education**	**Economic**	**Criminal**	**Mobile**
• Treatments and management of conditions ○ Current ▪ Health treatment ▪ Use of health services ○ Previous ▪ Health treatment ▪ Use of health services • Samples provided ○ Method ▪ Invasive ▪ Non-invasive ○ Type ▪ Blood ▪ Urine ▪ Hair ▪ Saliva ○ Storage of samples ▪ Rights ▪ Benefits and compensation ○ Tests and assessments ▪ Rights to results ▪ Length of test	Skills Authority ○ Educators ▪ Name ▪ Associated school/college/university			

Personal records				
Health	**Education**	**Economic**	**Criminal**	**Mobile**
▪ Location • Follow-up on health registration				

5.6.2.4 Information document

The final part of the model, information document, refers to the additional informational material which may accompany the consent form. Results of the analyses show that topics discussed included more information on the study and details regarding the withdrawal of consent. In the ALSPAC study for example, the accompanying information describes how data linkage works and how enduring the consent given is until withdrawn. The ALSPAC study is the only study to provide case studies of where record linkage has been used in the past e.g. I4C – researching childhood cancer. Results of the analyses also showed that in the ALSPAC and Life Study informational booklets, the benefits of participating in the study are outlined.

Furthermore, results of the analyses showed that in the BHPS, Scottish Health Survey, UK Biobank, Understanding Society and Life Study booklets, a description is provided of who may use your data. It was in the Life Study, Understanding Society, Born in Bradford and BHPS that data security/confidentiality are discussed.

Additionally, results of the analyses showed that there were different types of material available e.g. booklets, leaflets etc.; following the analysis, leaflets were identified for MCS and Understanding Society only. In the MCS, the first informational leaflet describes where data may be obtained whereas the second describes the study itself and the role of the child. In the Understanding Society study, one leaflet provided information on adding administrative health records and the other adding education records. The combined elements can be found in Table 5-6.

Table 5-6 Combined accompanying information booklet elements

Information document		
General information	**Participation**	**Record linkage**
• Study ○ Aims ○ Objectives ○ Funding bodies ○ Reviewers ○ Contact details • Confidentiality and security ○ Safeguards in place to protect participant confidentiality and data security	• Invitation process • Benefits and risks ○ Immediate ○ Future • Consent process ○ Coverage ○ Length of time ○ Withdrawal ▪ Levels of withdrawal • Visits ○ Prior ▪ Preparation ○ During ▪ Biological samples • Specify which ones • How will these be taken • Who will the samples be taken from ▪ Questionnaires to complete ○ Post ▪ Obtaining certain results • Other people ○ GP ○ School teachers • Expenses ○ travel	• Definition • How is it achieved • Case studies/examples • Data ○ Which records/registries will be linked to ○ How will the data be accessed • Subsequent research ○ Who will have access to the data ○ Getting to know results

5.6.2.5 Metadata management models

Having analysed each group of metadata elements, Table 5-7 shows these combined together. Based on this table, I then created five metadata management models using the information I had abstracted during the qualitative analyses. The models were created using a combination of conceptual modelling and object oriented techniques.

Figure 5-2 to Figure 5-5 Proposed person metadata management model present the four components of the metadata management model; the fifth - In Figure 5-6 I present a HTML report displaying a series of consent classes based on the elements from Table 5-7.

Table 5-7 Combined metadata elements of consent for record linkage in longitudinal studies

Personal records	People	Consent form	Information document
Health • Organisation ○ National Health Service ▪ The NHS Information Centre ▪ NHS Central Registrar ○ Department of Health ○ General Registration Office ○ Office for National Statistics • Healthcare professional ○ Primary care ▪ GP ○ Secondary care ○ Tertiary care • Clinical ○ Terminologies ▪ ICD-10 ▪ ICD for Oncology ▪ SNOMED-CT ▪ Read Codes ▪ DSM ▪ OPCS • Treatments and management of conditions ○ Current ▪ Health treatment ▪ Use of health services ○ Previous ▪ Health treatment	• Identifiers ○ NHS number ○ Passport number ○ other • Interviewee ○ Name ▪ Forename ▪ Surname ○ Address ○ Birth details ▪ Date ▪ Location ○ Ethnicity ○ Nationality ○ Disability ○ Contact details ○ Next of kin ○ Family ▪ Partner ▪ Dependents • Participant ○ Individual consenting for themselves ○ Parent/guardian on behalf of a child • Persons present ○ Date ○ Location	• Description/aims • Method of collection ○ CAPI ○ CASI • Undertakings ○ Declaration ○ Rights of participants • Organisations ○ Funding agencies ○ Universities ○ Governments ○ Archive • Questions/consent statements ○ Logic ○ Purpose ○ Potential Reponses ○ Codes and categories • Confirmatory information ○ Confirmation of understanding ○ Signature ○ Date	• Study ○ Aims ○ Objectives ○ Funding bodies ○ Reviewers ○ Contact details • Participation ○ Invitation process ○ Benefits and risks ▪ Immediate ▪ Future ○ Consent process ▪ Coverage ▪ Length of time ▪ Withdrawal • Levels of withdrawal ○ Visits ▪ Prior • Preparation ▪ During • Biological samples ○ Specify which ones ○ How will these be taken ○ Who will the samples be

Personal records	People	Consent form	Information document
▪ Use of health services ○ Samples provided ▪ Method • Invasive • Non-invasive ▪ Type • Blood • Urine • Hair • Saliva ▪ Storage of samples ▪ Rights ▪ Benefits and compensation ○ Tests and assessments ▪ Rights to results ▪ Length of test ▪ Location • Follow-up on health registration Education • Type ○ School records ○ Further education ○ Higher education • Provider ○ Organisation • Department of Education • The Data Service • Department for Business,	• Interviewer ○ Name ▪ Forename ▪ Surname • Witnesses ○ Name ▪ Forename ▪ Surname	○ Full name	taken from • Questionnaires to complete ▪ Post • Obtaining certain results ○ Other people ▪ GP ▪ School teachers ○ Expenses ▪ travel • Record linkage ○ Definition ○ How is it achieved ○ Case studies/examples ○ Data ▪ Which records/registries will be linked to ▪ How will the data be accessed • Subsequent research ○ Who will have access to the data ○ Getting to know results • Confidentiality and security ○ What safeguards are in place to protect participant confidentiality and data security

Personal records	People	Consent form	Information document
Innovation and Skills • Universities and Colleges Admission Service (UCAS) • Higher Education Statistics Agency • Department for Children, Schools and families • Department for Children, Education, Lifelong Learning, and Skills • Government Education Directorate • Department of Education/Education and Skills Authority ▪ Location ○ Educators ▪ Name ▪ Associated school/college/university Criminal • Organisations ○ Ministry of Justice • Records ○ Official cautions ○ Convictions Work and employment • Organisations ○ Department for Work and Pensions ○ HM Revenue and Customs			

Personal records	People	Consent form	Information document
• Records ○ Salary ○ National insurance contributions ○ Tax ○ Savings ○ Benefits ○ Pensions Mobile • Past • Current • Future			

Figure 5-2 Proposed information document metadata management model

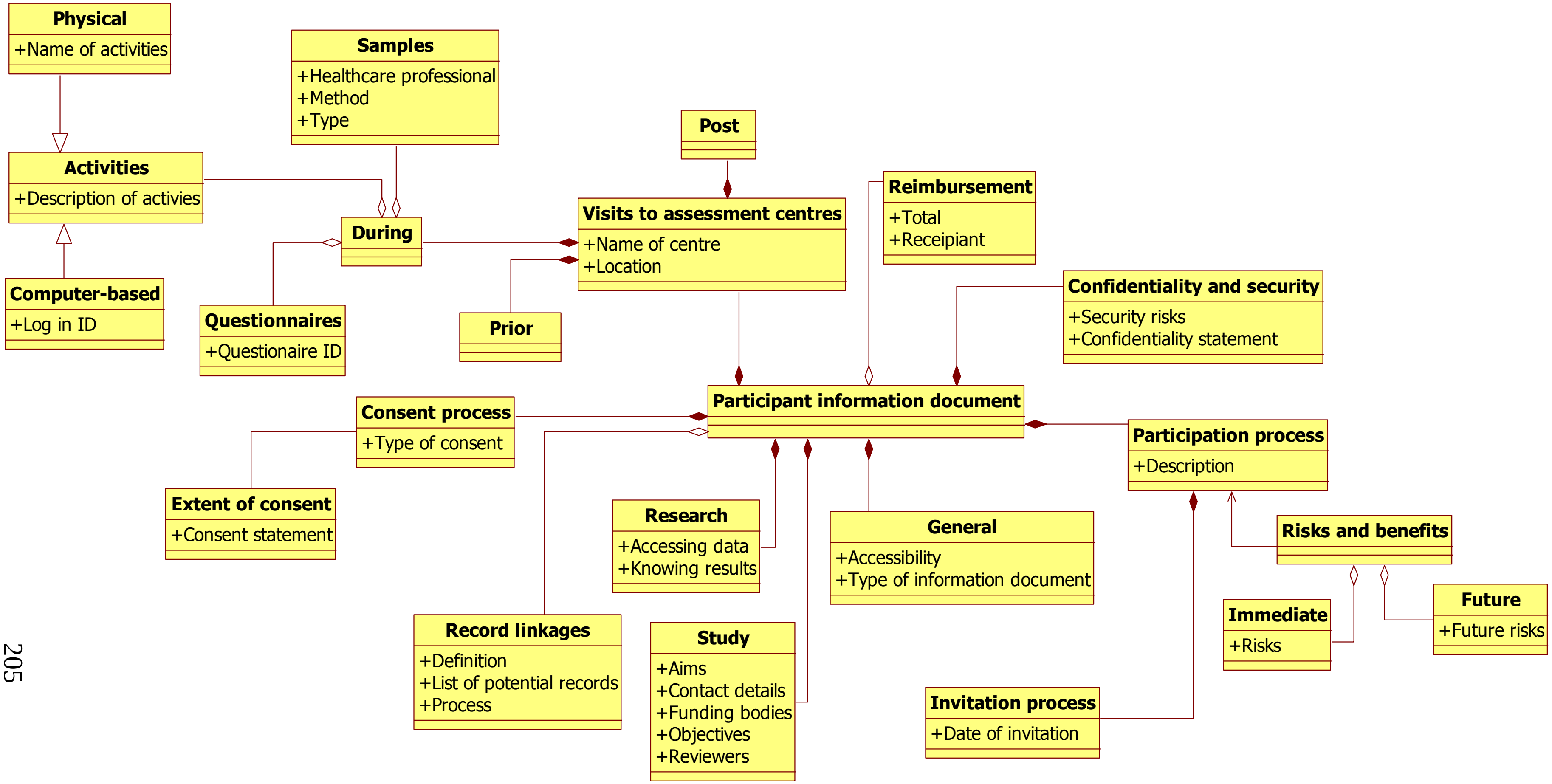


Figure 5-3 Proposed consent form metadata management model

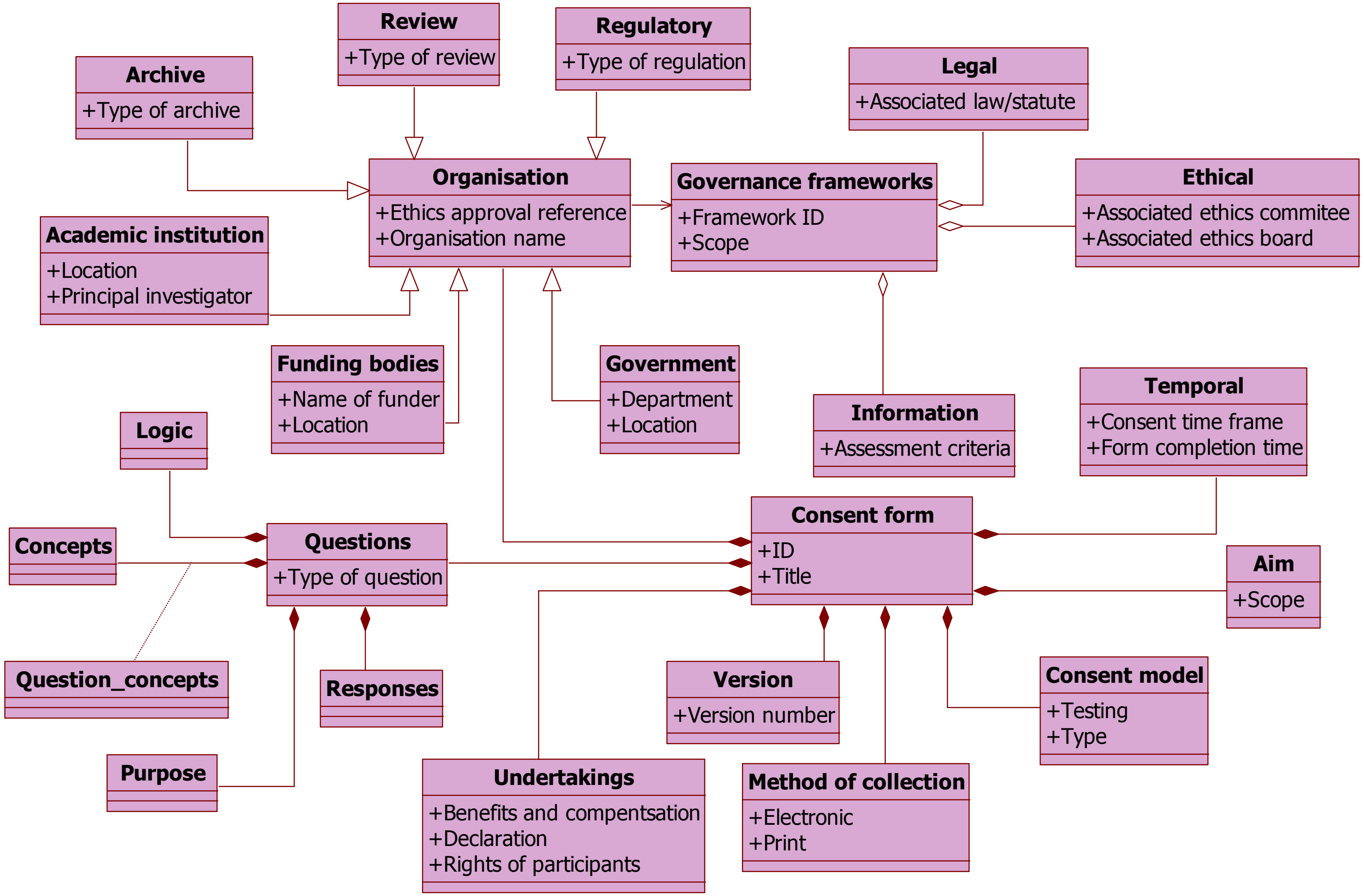


Figure 5-4 Proposed records metadata management model

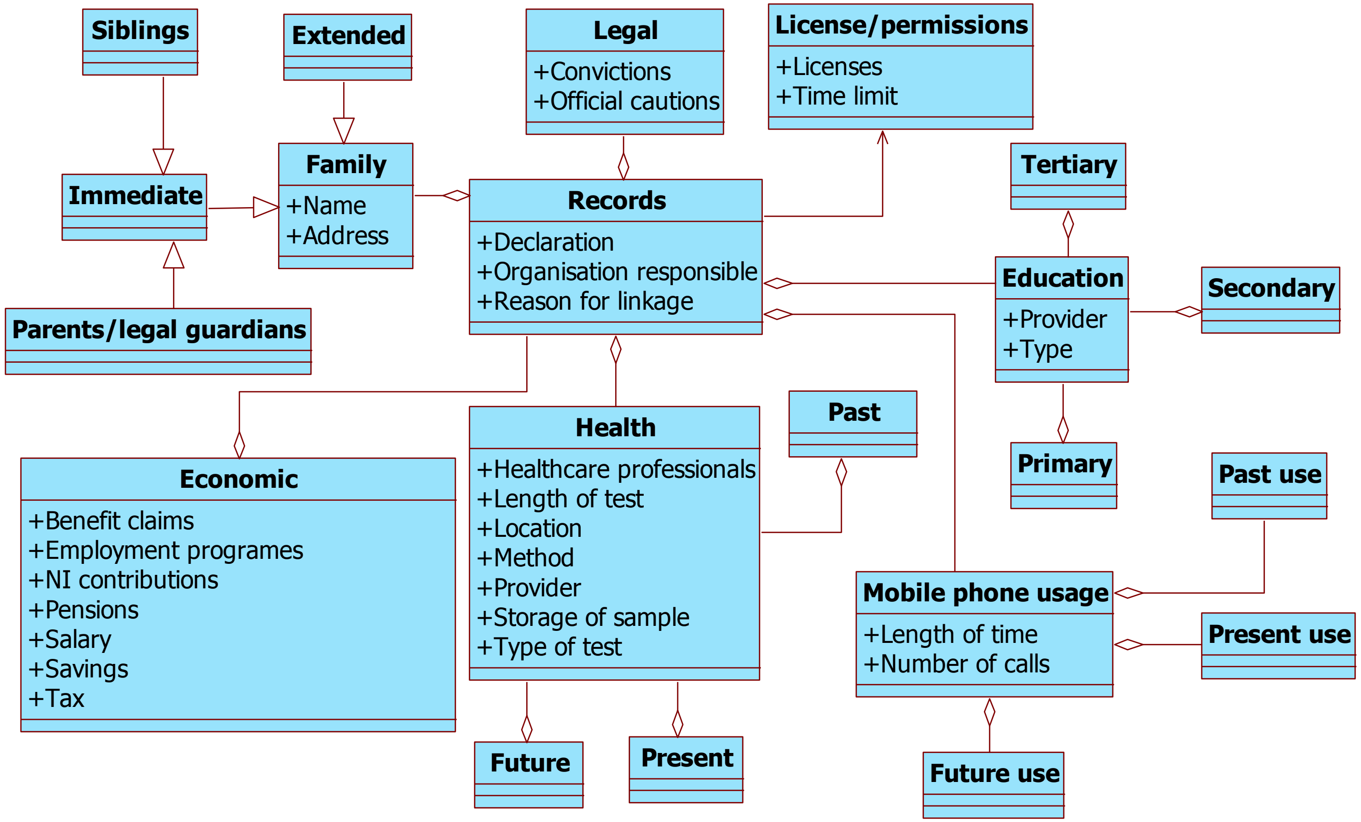


Figure 5-5 Proposed person metadata management model

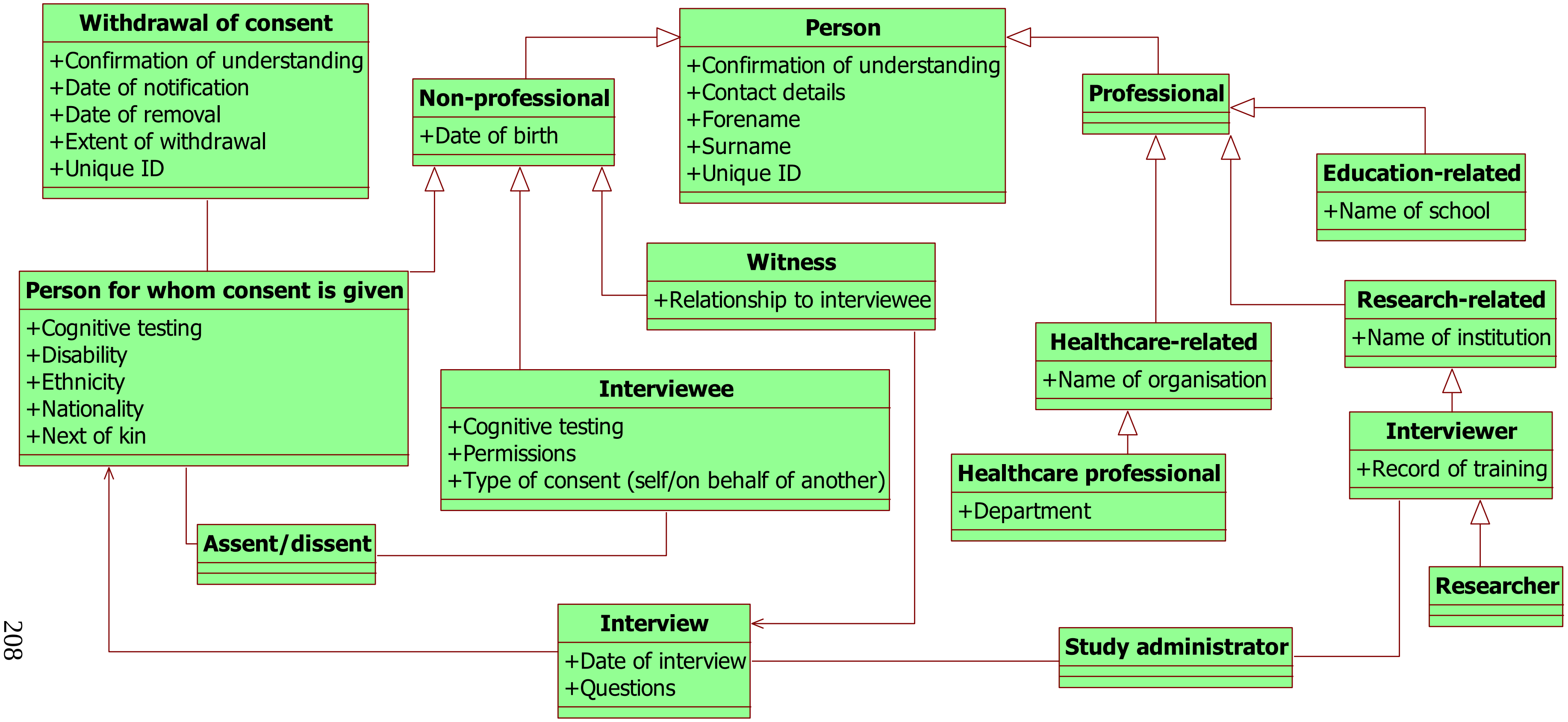
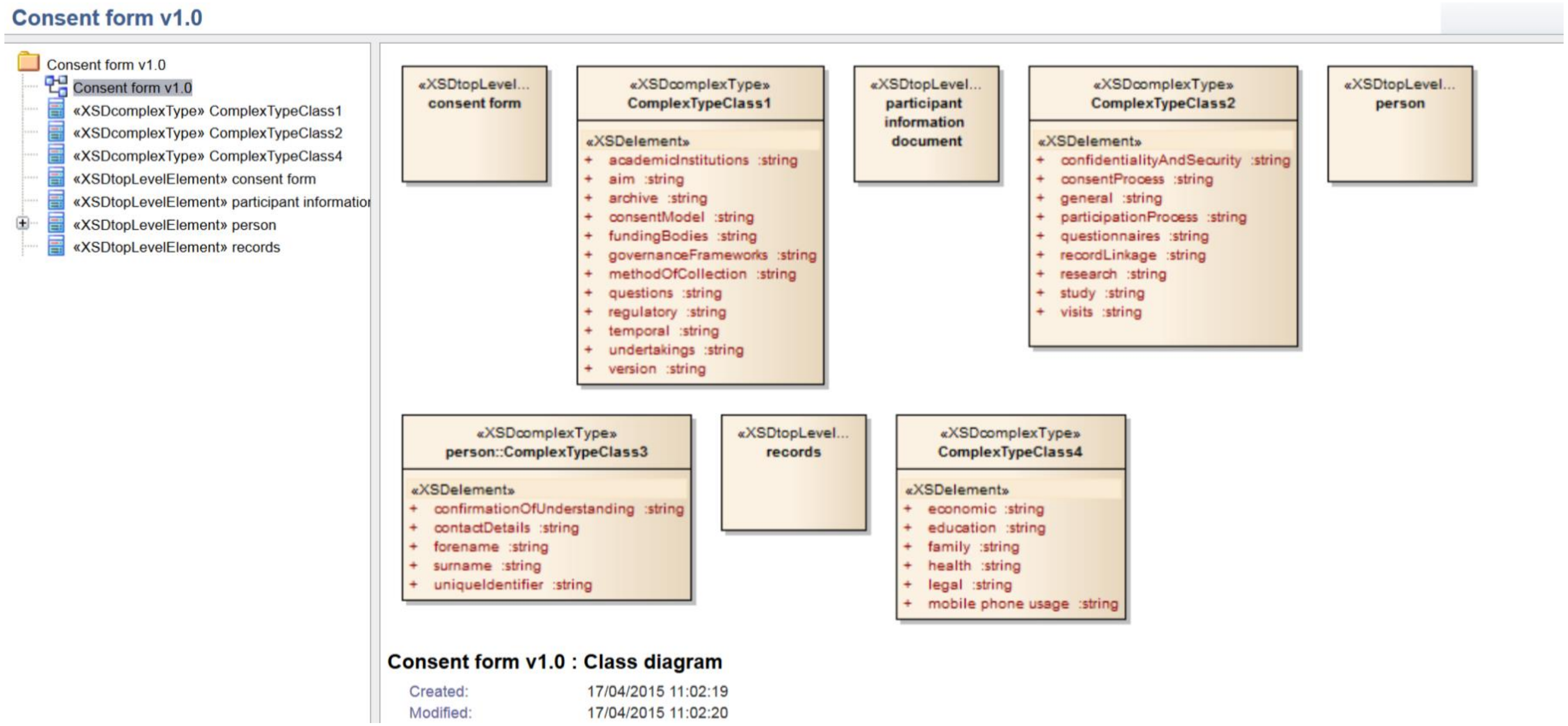


Figure 5-6 HTML report based on combined metadata management model



Consent form v1.0

- Consent form v1.0
  - Consent form v1.0
  - «XSDcomplexType» ComplexTypeClass1
  - «XSDcomplexType» ComplexTypeClass2
  - «XSDcomplexType» ComplexTypeClass4
  - «XSDtopLevelElement» consent form
  - «XSDtopLevelElement» participant informatio
  - «XSDtopLevelElement» person
  - «XSDtopLevelElement» records

Consent form v1.0 : Class diagram

Created:	17/04/2015 11:02:19
Modified:	17/04/2015 11:02:20

5.6.3 DDI 3.2 critical evaluation

I mapped the consent metadata elements to the Data Documentation Initiative 3.2 (DDI 3.2) elements to determine the extent to which consent for record linkage in longitudinal studies may be recorded using the prevailing existing metadata standard. The evaluation had four components: a) personal records; b) people; c) consent form; and d) information document. The cross-walks test how well consent metadata elements identified in Table 5-7 can be mapped directly to elements in DDI 3.2.

5.6.3.1 People

Results of the first mapping analysis (Supplementary Table 24) showed that general metadata elements such as full name, location and nationality can be mapped directly to DDI 3.2 using the ‘FullName’, ‘LocationName’ and ‘Country’ elements respectively. It was also possible to map to DDI 3.2 directly for other elements such as contact details (‘TelephoneNumber’ and ‘Email’) and date (‘Date’). By using these elements in DDI 3.2, in addition to others, descriptions of the people involved in the consenting process can be created relatively quickly.

However, a fundamental aspect of consenting process associated with longitudinal studies is the possibility for the interviewee to consent on behalf of another person; for example, a mother consenting on behalf of her infant. Recording this information in DDI 3.2 is a more complex process; I was unable to locate an element designated to recording this information. DDI 3.2 does not contain an element to record this kind of information. Currently, stakeholders could use the ‘Note’ element along with another object to record this information. Though this solution works in the short term, as the infant matures, and they become more able to assent/dissent, DDI 3.2 lacks the necessary mechanisms to record this metadata. Therefore, additional elements (e.g.for whom consent is given) are needed in DDI to enable the recording of these critical pieces of metadata.

5.6.3.2 Consent form

Results of the second mapping analysis (Supplementary Table 25) focused on identifying and mapping to DDI 3.2 elements to record the consent form's composition. This process was fairly straightforward and the majority of elements could be mapped directly. This is because DDI 3.2 is designed to capture survey instruments and this is what a consent form essentially is. In places, text can be recorded using 'ResponseText'; or stakeholders could select a response from a predetermined, named list and use a combination of, 'CodeList', 'CodeListName' and 'CodeListReference' DDI elements.

Nevertheless, DDI 3.2 again lacks the elements needed to record details specific to consent to record linkage in longitudinal studies. For example, 'Undertakings' and 'Confirmatory information' could not be mapped directly to a DDI 3.2 element. Instead, stakeholders may use 'Note' element to hold the necessary information which can then be attached to another maintainable object. It is in these areas in particular that the standard fails to provide the necessary mechanisms to record consent for record linkage using metadata elements effectively. Therefore, extensions are needed to the DDI 3.2 standard to enable to the recording of the different aspects of consent forms for record linkage.

5.6.3.3 Personal records

Results of the third mapping analysis (Supplementary Table 26) indicated that a number of different elements associated with personal records can be recorded using DDI 3.2. For example, I could group together organizations and assign a group name. Here, the element 'CodeListGroup' can be used to specify the name of the group; while 'CodeList' will enable stakeholders to record the possible codes such as ICD-10. The DDI has built-in mechanisms to successfully record code lists and categories so the process of providing links between these elements was relatively straightforward. Other areas in which the DDI provided the necessary mechanisms to map directly included location - 'LocationName'.

However, results of this cross-walk analysis demonstrated that a direct link from elements associated with, 'Treatments and management of conditions' and 'Tests and assessments' to DDI 3.2 was not possible. Having access to this kind of metadata in a standardised and simplified format is key to supporting stakeholders in determining use of health services (part of current treatment and management of conditions) and rights to results (part of tests and assessments). This lack of opportunity for mapping between the two could be because this level of functionality has not been previously needed in DDI and so does not contain mechanisms to record this metadata.

5.6.3.4 Informational document

Results of the last mapping analysis (Supplementary Table 27) indicated that DDI 3.2 can successfully record the study and its objectives can be mapped directly using the ‘Citation’ element. Other elements such as funding bodies can also be mapped directly using the, ‘FundingInformation’ element. Another advantage of using DDI 3.2 is that lifecycle events, such as a participant withdrawing their consent can be recorded, and mapped with relative ease. Here, a combination of ‘EventType’ and ‘LifecycleEvent’ would enable stakeholders to record this information in a clear and standardised manner. This is very important as the boundaries of consent have a direct impact on researchers wanting to use certain research data. DDI 3.2 also enables stakeholders to create multiple lifecycle events. This can be used to the stakeholders’ advantage as they can record any additional relevant events in a systematic and robust manner.

Although, stakeholders are again restricted in recording details specific to consent for record linkage. For example, there are no elements in DDI 3.2 which can be mapped directly enabling the recording and use of biological samples and how these will be taken. Being able to record, and have access to, this kind of metadata is important to informing potential secondary users of which biological samples could be used as part of their further analyses of the research data. Furthermore, having access to this kind of metadata can also help other stakeholders, such as potential participants understand what could be requested of them should they partake in the study. This metadata is critical to informing these and other stakeholders and ensuring potential participants are even more informed prior to engaging with the study.

5.6.4 Evaluation of consent for record linkage metadata management model

5.6.4.1 Person

I applied the model to the first test case, ELSA, and found that the model provided the metadata elements needed to record information about the interviewee and the interviewer. However, the model lacks elements to record when the consent form was completed; this indicated that the model needed revising to enable recording of this information. In terms of recording confirmation of understanding, I was able to use the ‘Confirmation of understanding’ attribute as part of the ‘Person’ element. This is a key aspect of the model as the need for a standardized approach to recording this kind of information is needed and this section of the model address this current unmet need. I then applied the model to the second test case, CLSA. I found that the metadata relating to the people involved could be recorded using the ‘Person’ element and the child elements, ‘Non-professional’ and ‘Professional’. In having two separate child elements, I was able to reduce the number of repeating attributes in the model through use of inheritance; whilst, enabling ourselves to distinguish between the different types of people involved in the consenting process – ‘professional’ e.g. principal investigator and ‘non-professional’ e.g. interviewee. Following application of the model to the second test case, I decided to create a new element, ‘date of completion’, and that this should be moved, and joined to, the ‘consent form’ element. In doing so, I was able to group together all the elements relating to the consent form itself to help users better navigate through the model. The final test case used to evaluate the model was the LSAC study. Having iteratively applied the model to the previous test cases and made changes, I was able to record information such as ‘confirmation of understanding’ successfully.

5.6.4.2 Consent form

I firstly applied the model to ELSA to evaluate how well our model could record information relating to the consent form’s composition. I found that use of elements such as ‘Academic institution’ enabled us to record detailed information, particularly since this element inherits the 'Ethics approval reference' and 'Organization name' attributes from the 'Organization' element. In using object-oriented modelling techniques, I was able to harness the advantages using a typically computer science technique can bring to life sciences research. Following this initial test, I identified areas for improvement. For example, the ELSA consent form contains a set of instructions detailing what to do with the completed form - one copy is retained by the participant and the other is returned to the office. The model does not contain an element to record this information and so an additional element was needed. I altered our model accordingly, by adding 'Instructions for next steps', and applied the revised version to the second test case. Results of the second test, using the CLSA consent form, showed us that by adding the new element, I was then able to record more detailed information about this aspect of the consent process. Results of the second test also demonstrated that by harnessing the element such as 'Questions', as composed of 'Logic', 'Responses', 'Purpose', forming a part of the 'Data collection', I was able to record the questions and question logic of the consent form in detail. This is important as being able to identify and record the minutia around this aspect of the consent process can potentially give stakeholders greater support in determining the scope of consent. The model was then reapplied to the final test case, LSAC. Here I was able to test how well the model could, for example, record introductory information. Recording this information involved use of the 'Aim' and 'Undertakings' elements which are a part of the 'General' elements. The next step was to test the records component of the model.

5.6.4.3 Personal records

The model contains six different types of personal record: 'Economic', 'Education', 'Legal', 'Family', 'Mobile phone Usage', and 'Health'. I applied the model to the first test case, ELSA, and found I was able to record previous hospital visits and treatments through the 'Health' element and one of its child elements, 'Past'. In having a child element 'Past', in addition to two others, 'Present' and 'Future' I was able distinguish between these different events. In terms of recording economic information for example, I used the ‘Economic’ elements with attributes: ‘Benefits claims’, ‘NI contributions’ and ‘Tax’. However, to enhance the model further, I converted these attributes into separate elements which, when combined, create the ‘Economic’ element. I made this change to simplify the recording of this kind of information; in terms of cardinality, there would be no restrictions on the number of times elements can be used. I then applied the model to the second test case and were able to record health related information using the existing ‘Health’ element’. For the third test case, the LSAC study, I again used the ‘Health’ element but also used the ‘Persons’ element in addition to sufficiently record this information. Results of these tests demonstrate that the model can record information about consent to use of personal records in longitudinal studies.

5.6.4.4 Informational document

The last section of the model I evaluated was information document. Results of the application to the ELSA study consent form demonstrated that to record different kinds of informational documents, the parent element, ‘Participant Information document’ and the child element ‘General’ needed to be combined to form a new element, ‘Informational document’ with ‘accessibility’, ‘audience’, and ‘type as attributes’. This enabled the recording of, and differentiation between, the different types of informational document. However, I was not able to locate additional informational material online for ELSA; I also experienced this problem for the CLSA study. Therefore, I decided to proceed to the third test case, LSAC, to continue testing this aspect of the model. During the final test, I was able to use the newly created, ‘Informational document’ to specify the document type – in this case it was the corresponding information sheet. In making this change to the model, I reduced the total number of elements whilst increasing scope to record metadata relating to the different kinds of informational document. This is an advantage of using a formalized modelling technique as changes were made quickly without impacting the rest of the model. I then recorded the description of the study using the ‘Study’ element with attributes, ‘Aims’, ‘Contact details’, ‘Funding bodies’, ‘Objectives’ and ‘Reviewers’.

Figure 5-7 Evaluated metadata management model for records

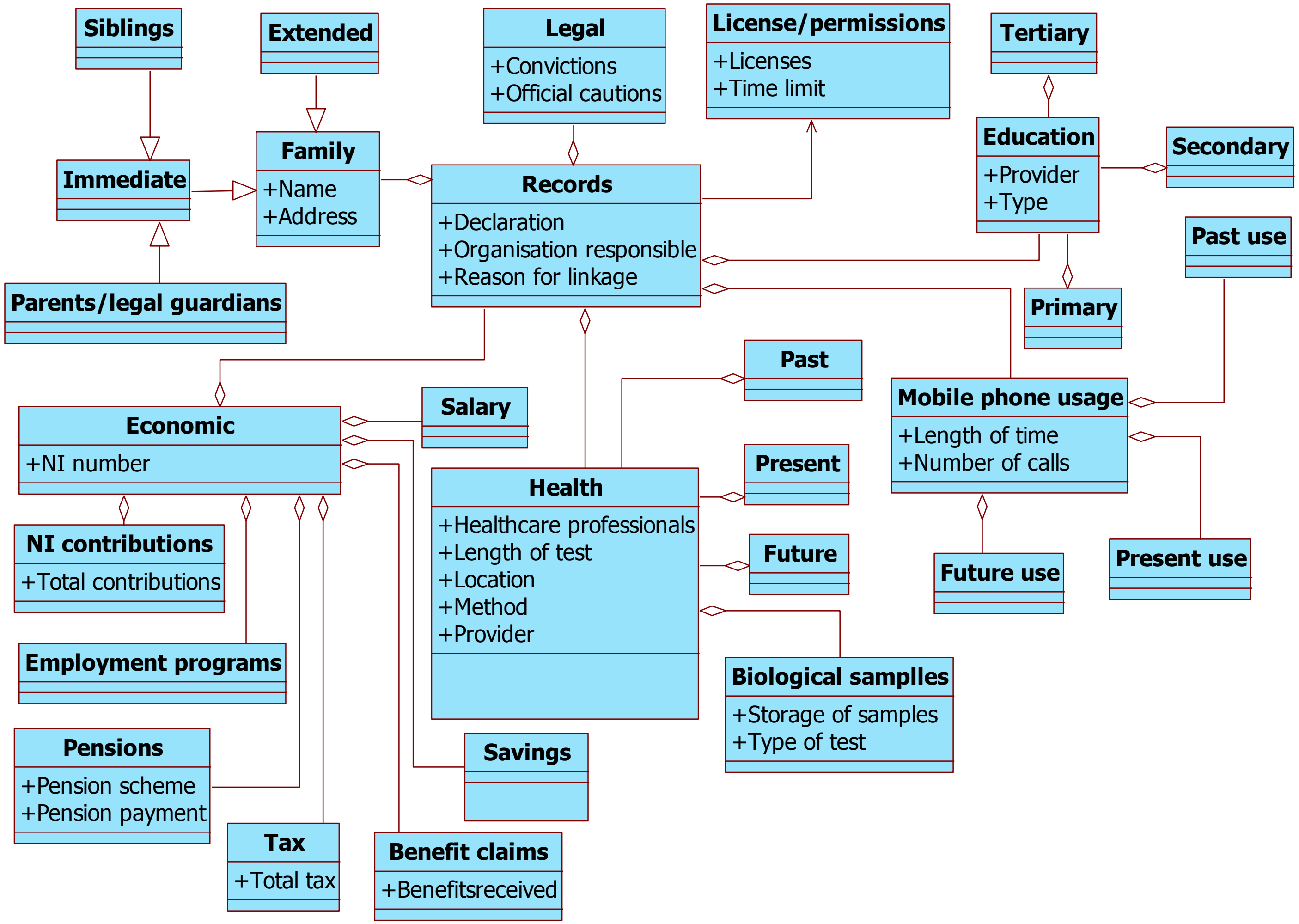


Figure 5-8 Evaluated metadata management model for person

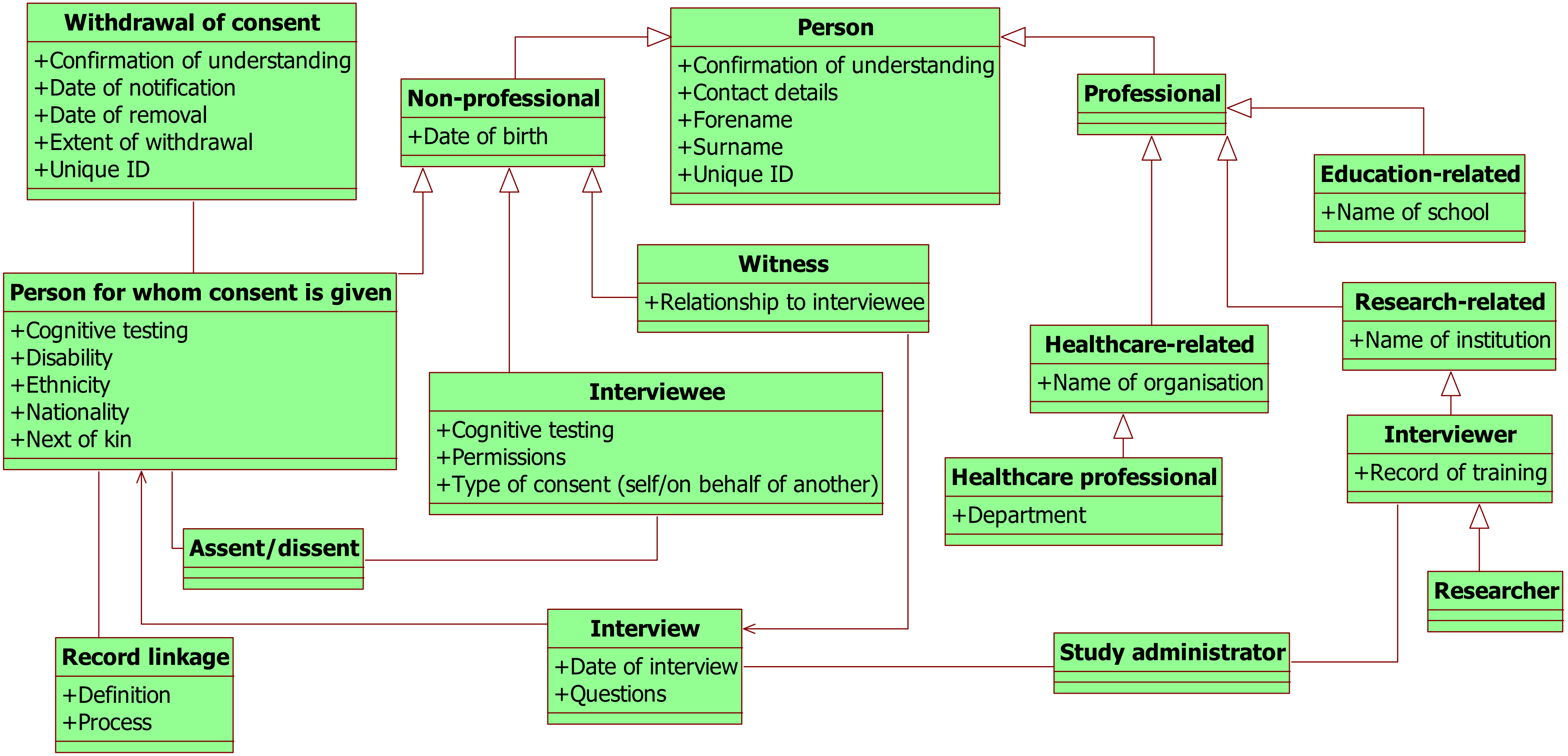


Figure 5-9 Evaluated metadata management model for consent form

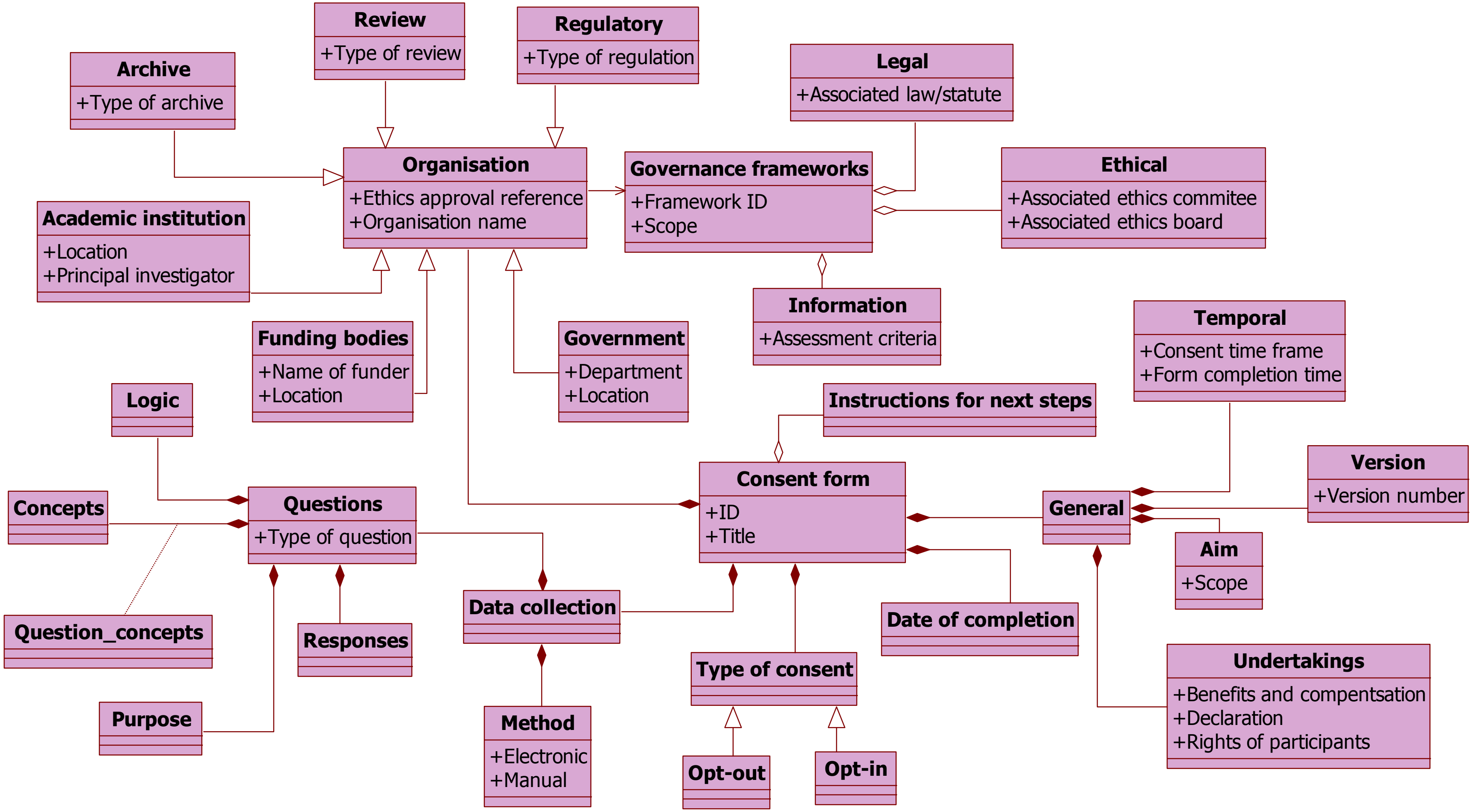
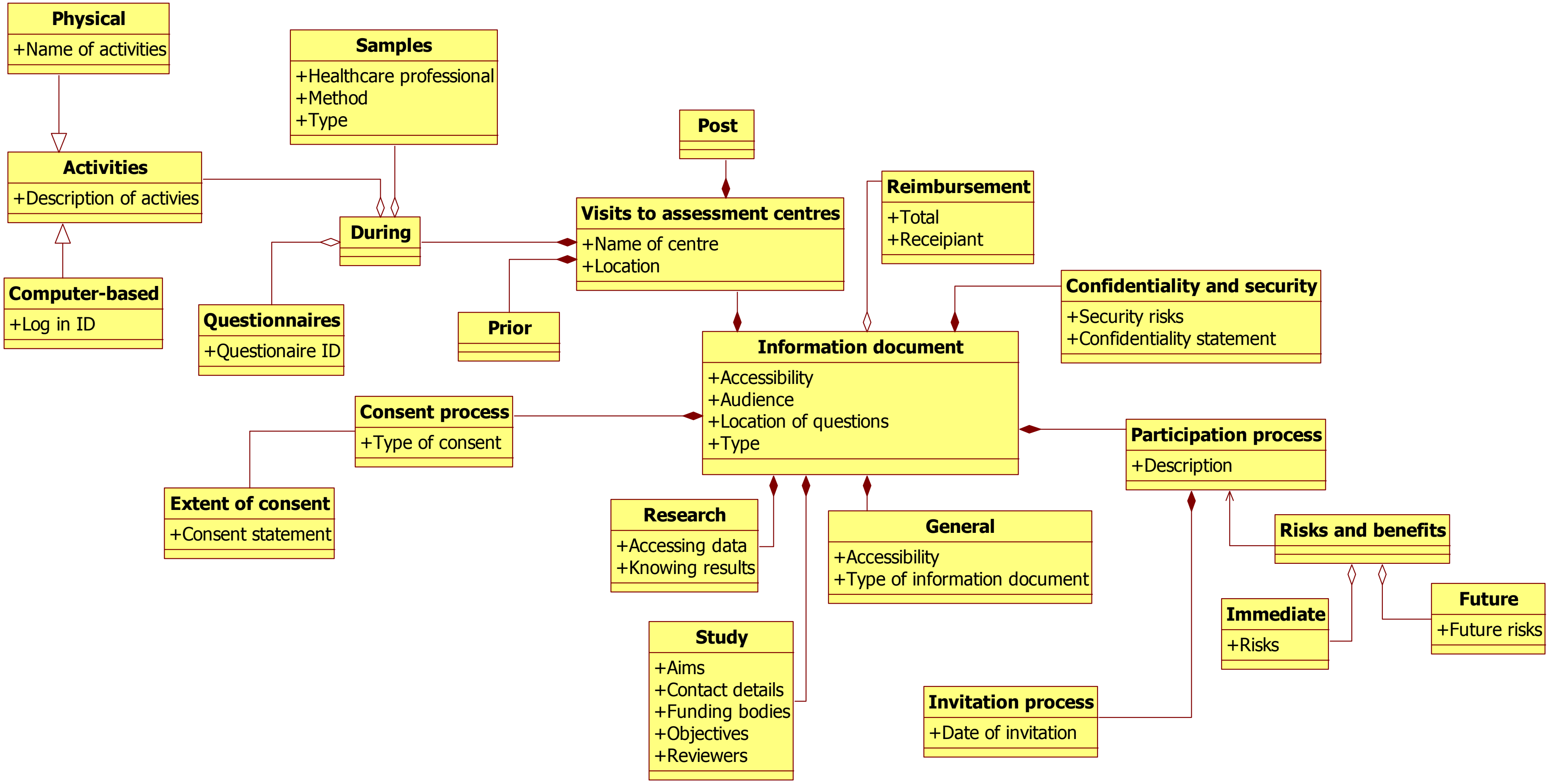


Figure 5-10 Evaluated metadata management model for information document



5.7 Discussion

5.7.1 Literature review

A total of 61 manuscripts were identified and reviewed but none were identified focusing on the development of models to record consent for record linkage in longitudinal studies. The literature review suggested that research into the development of these models is limited.

Following the review I was able to begin gathering requirements for the metadata management models. The literature I identified provided the contextual information needed to better my understanding of the overall consent process. Findings from the literature review demonstrated the need for the development of metadata management models to facilitate the standardised capture of consent metadata elements.

The strength of the literature review lies in the way it was systematically conducted using the PRISMA checklist. Both computer science and biomedical databases were used to increase the potential for sourcing literature for review. A potential weakness in the literature review however lies in the combination of search terms. The terms used were specific to public health and epidemiological research and not readily used in other domains such as computer science. Therefore, possible differences in use of controlled vocabularies may have resulted in literature being missed.

The next step was to create and evaluate a novel model that uses metadata elements to record this kind of information.

5.7.2 Metadata model design and development

Following analysis of the consent forms, a combined list of elements for each component of the model was established. These components were: a) personal records; b) people; c) consent form; and d) information document. Findings from each of the individual analyses were combined into a single table which presents all the metadata elements for consent for record linkage.

People: it was identified that for the ALSPAC and UK Biobank studies, consent could only be given from the participant themselves. As the consent forms reviewed are not necessarily the only forms associated with these studies, it is possible that other forms are age/relationship restricted and that in this instance that differentiation was not evident. Therefore, in terms of informing the development of the model, a mechanism is needed through which this distinction may be made.

Furthermore, developing a simplistic yet highly detailed model required the use of inheritance to create a parent class, ‘Person’, in which commonly identified elements such as full name appearing as attributes. This parent class also contains attributes such as, ‘Confirmation of understanding’, ‘Forename’, ‘Surname’ and ‘Unique identifier’ which are inherited by the child classes, ‘Professional’ and ‘Non-professional’. By having these two child classes, stakeholders can differentiate between those working for the study – professionals, and those who are (potentially) a part of the study – non-professionals. The use of inheritance here also serves to further define the different types of professionals e.g. ‘healthcare-related’ such as GPs, ‘research-related’ such as principle investigators and ‘education-related’ such as teachers. This principle will also be applied to the non-professionals where elements such as, ‘interviewee’ and ‘witness’ will be included.

Moreover, the model needs to differentiate between the interviewee and person for whom consent is being given. This is because a participant may be consenting for themselves, as in the UK Biobank study, or be consenting on behalf of another, e.g. in the Life Study. Being able to record clearly for whom consent is given is key as a parent/guardian may be consenting on behalf of a child; a situation which may change as the child matures, autonomy increases and becomes more able to assent/dissent. Therefore, there are two separate elements in the model, ‘Interviewee’ and ‘Person for who consent is given’. By adding an associated element, ‘Withdrawal of consent’ with an attribute, ‘Withdrawal of consent’, this will facilitate the recording of metadata such as date of notification, extent of withdrawal and confirmation of understanding. This is important as there are instances whereby if data have been anonymised, aggregated and shared with other parties’, destroying this information is not always feasible. By having this attribute, the interviewee can confirm whether this, along with their right to withdraw consent has been understood.

Consent form: I noticed that all the consent forms followed a similar structure – introduction, consent questions/statements and confirmation of understanding including dates and signature(s). In terms of model development, principles such as composition enabled me to build parts of the model individually, to then demonstrate how different elements can then combine to form something else. For example, the responses, logic and purpose combine to form questions; these questions can then be combined with other elements to form the consent form. By adopting a modular approach to model development, this helped me to build the model systematically and identify where aggregation may also be applicable improve the model's design.

Additionally, the studies (except for UK Biobank) stated on the consent form from where data may be obtained. Given the potential for information to be sourced from multiple, different data holders, a mechanism was needed in the model to record where data have been received and the scope/reason for the data/data linkage. Therefore, I used inheritance to define the relationship between the parent and more generalised class, 'Organisation' and the more specialised, child classes such as 'academic institutions', 'governments' and 'regulatory bodies'. These child classes can inherit common attributes contained in the parent class, such as organisation name, and use these. Consequently, stakeholders will potentially be better able to distinguish between the different types of organisations and from where records were sourced.

Personal records, I noticed that age plays a role in deciding which consent form to complete. For example, in the Understanding Society study for example, the distinction is made between children aged 0-15years and those aged 16-24 years; consequently, two separate forms are used. This is so the parent/guardian can consent on behalf of a child aged 0-15 whilst those aged 16-24 may consent for individually. Being able to record this distinction in the model is important as it reflects the increasing autonomy children have throughout the course of a longitudinal study and their rights to assent and/ dissent. Therefore, the element ‘Records’ had to have an attribute ‘Declaration’ where this information can be specified.

Furthermore, I decided to use a combination of aggregation and inheritance to ensure the model is built as efficiently as possible, enabling stakeholders to record the different types of records. By using aggregation, stakeholders can select which elements they need e.g. ‘Education’, ‘Health’, Economic’ etc. whilst still being able to indicate that these elements combine to form the ‘Records’ element. Inheritance will be used to indicate how certain elements are connected. For example, the element ‘Parents/Legal Guardians’ will inherit attributes and methods from the ‘Immediate’ element which in turn will inherit attributes and methods from the ‘Family’ element. I will use aggregation to connect the ‘Family’ element to the ‘Records’ element.

Results of the analysis also showed that elements such as ‘Education’ and ‘Health’ are composed of other elements. For example, education records may be sourced from all three sectors of education. Therefore, the model will have the elements, ‘Primary’, ‘Secondary’ and ‘Tertiary’ which will aggregate together to form the ‘Education’ element. The ‘Health’ element will be aggregated of, ‘Past’, ‘Present’ and ‘Future’. This enables stakeholders to record the different types of health data available to them. This is important to the (on-going) development of the model as in studies such as the Health Survey for England and Scottish Health Survey request to follow up is made.

Additionally, results of the analysis showed that consent to access and use economic records was requested from the, Life Study, BHPS, and ALSPAC. To maintain simplicity in the model and yet allow stakeholders to record more low level metadata, the ‘Economic’ element has attributes such as ‘National insurance contributions’, ‘Benefit claims’ and ‘Tax’. By giving the ‘Economic’ element these attributes, detailed descriptions may be recorded in a standardised manner.

Informational documents: In this component of the analysis, the elements identified could be categorised into three groups: a) general information; b) participation; and c) record linkage. In terms of informing development of the model, I used composition to show that the ‘Participant information document' element is made up of several other elements. For example, the 'Study' element will have attributes such 'Aims' and 'Research objectives'; whilst, the 'Participation process' element will be associated with a 'Risks and benefits' element – another common theme in the information documents.

Furthermore, several of the studies (Life Study, Understanding Society, Born in Bradford and BHPS) provided a description relating to data security/confidentiality. In terms of informing development of the model, an element was needed to enable the recording of this metadata and specify (if necessary) from where this metadata was sourced. Additionally, for studies such as the MCS and Understanding Society, several informational leaflets were identified and reviewed. In terms of informing development and use of the model, stakeholders must be able to record multiple items; hence, any associated cardinality/multiplicity is reflective of this. Table 5-7 shows the finalised complete set of elements for consent to record linkage metadata in longitudinal studies.

By adopting an object oriented approach to the design of the model, use of inheritance in particular, enabled multiple elements to share attributes. By using inheritance I was able to reduce repeating elements in the model, potentially increasing its efficiency. For example, in the consent form section using the third test case, in having a generalized parent element of 'Organization', attributes such as 'Organization name' are inherited by every child element helping to produce a simplified model capable of recording low level detail. The sharing of attributes can potentially decrease the number of repeating, thus redundant, elements.

The statements on the consent forms were recorded using the 'Questions' element. The use of aggregation enabled us to specify the elements needed to compose metadata relating to the questions, or in this case, the three statements. This is an advantage of using object oriented modelling techniques to design and develop the model. As part of section D, an example is provided describing potential information that may be accessed. This may be captured using the 'Research' element as part of the

‘Informational document’ element. Use of aggregation here enabled a detailed and structured description to be constructed.

5.7.3 DDI 3.2 critical evaluation

Results of the evaluation indicated that whilst many DDI 3.2 elements may be harnessed to create standardised descriptions of consent for record linkage in longitudinal studies, the standard lacks the mechanisms needed to record low level metadata specific to epidemiological and public health research. For example, it was challenging to identify and select which elements could be used individually or together to record the interviewee’s decisions and the reason(s) why. It is having access to these details which could potentially better support researchers in undertaking analyses using record linkage.

Nevertheless, the strength of DDI sits very much in the opportunity for stakeholders to package together standardised instances of metadata in an interoperable format (XML) which can then be published as a ‘StudyUnit’. These instances may be entered into inter/national catalogues where they can be actively maintained. Subsequently, stakeholders such as potential secondary users, members of the public, and in addition to others, may view these metadata records to better inform themselves of past and current longitudinal studies. Consequently, stakeholders can potentially have a greater understanding of what could be achieved if access to the data and/or biological samples was granted.

This evaluation served as additional confirmation for the need to develop a novel model to record consent for record linkage using metadata elements. It also demonstrated the areas in which greater focus in needed to improve the recording of low level metadata relating to longitudinal studies in epidemiological and public research settings.

5.7.4 Metadata model evaluation

Currently, identifying metadata elements involves a qualitative analysis of the consent form and there is the potential for bias given the potential subjectivity of the process. This is a potential weakness of the model. Further, there is a limit to which these analyses may be conducted manually particularly when on a much greater scale. Consequently, this impacts the extent to which it is ready for use in real world applications. Ideally, the entire process would be automated and identification of concepts would be through use of a predefined concept list, possibly structured using some kind of ontology, from which concepts may be selected and assigned. In having a fully automated process, this could potentially reduce human error and scope for bias. Furthermore, information documents could not be found online for ELSA and CLSA studies. This is a weakness in our approach to evaluating the model as this section of the model was not tested to the same extent as the other three sections.

To enhance the model further, the 'type of test' and storage of sample' attributes of the 'health' elements should be removed and placed in a new element entitled 'biological samples'. This is because having a separate element for this information widens the scope for further extension and enables additional, element-specific attributes to be added such as name and site of labs. This could also potentially improve the extent to which dynamic consent may be captured as changes in the model would facilitate the recording of more low level detail. In the records section, extending the health element to include a separate element for biological samples would also enhance the model further.

5.8 Conclusions

Longitudinal studies are critical to investigating the aetiology of disease and its impact across the life course. Performing recording linkage using personal records accessed through longitudinal studies can create datasets with even greater complexities which lend themselves well to thorough analyses. Hence, recording consent for record linkage in longitudinal studies in a standardized and robust manner is vital to supporting the record linkage process.

I created and evaluated a novel metadata model to record consent for record linkage in longitudinal studies using metadata elements. This addresses the current unmet need for architectures to support the systematic recording of such information.

Next steps include engaging with stakeholders to further evaluate the model and discuss how the model may form a part of stakeholders’ work routines. The longer term goal is to integrate this model into a tool which will help to quicken the recording process through automated or semi-automated processes.

5.9 Summary of major findings

5.9.1 Literature review

A total of 61 publications were fully reviewed of which none described the development of models to record consent to record linkage metadata. This underlines the unmet need for approaches to support the systematic recording of such information.

5.9.2 Metadata model design and development

A total of 30 consent forms from nine longitudinal studies were qualitatively analysed. The analysis involved identifying the metadata elements belonging to four categories: a) people involved in the consent process; b) composition of the consent form; c) personal records; and d) additional informational document. Individual analyses were conducted before these were combined and repeating elements were removed.

5.9.3 DDI 3.2 critical analysis

I mapped the metadata management models to test how well the consent metadata elements could be mapped directly to elements in DDI 3.2. Results of the analyses showed that generalised metadata elements can be mapped directly whilst others may be linked using a combination of elements including ‘Note’. The results demonstrated that DDI 3.2 lacks the mechanisms needed to produce detailed descriptions of the consenting process indicating the need for a novel model to do so.

5.9.4 Model evaluation

A total of five consent metadata management models were created with a corresponding XML schema. The models were built using object oriented modelling techniques. By way of evaluation, the metadata management models iteratively applied to the consent to record linkage metadata for the ELSA (wave 6), CLSA study and LSAC. Following each application, changes for improvement were identified and corrective action taken.

5.10 Chapter summary

This chapter focused on exploring the current state of art in capturing metadata for record linkage metadata in longitudinal studies, identifying the key elements of consent metadata, creating a novel metadata model to capture these and evaluating the model.

I performed a systematic literature and conducted a qualitative analysis of 30 consent forms sourced from nine different longitudinal studies. The literature review did not identify any literature focusing on the development of consent models for record linkage metadata within the context of longitudinal studies. Results of literature review signal the lack in formalised methods to record this type of metadata and the need to develop a model to address this gap in knowledge.

Results of the consent form analysis show metadata elements for consent to record linkage can be categorised into four components: a) personal records; b) people; c) consent form; and d) information document. The elements of these components were then cross-walked with the elements in DDI 3.2.

Once the qualitative analyses and DDI 3.2 critical analysis were completed, a novel metadata management model was created to capture the metadata elements associated with consent to record linkage. I evaluated this model by applying it to a series of test cases and identifying the strengths and weaknesses. Changes were made to the model iteratively until it had been fully evaluated. The metadata management model can now be used to assist the capture of consent for record linkage metadata in longitudinal studies. It can also form the basis of development work on creating a tool to help automate this process.

Chapter 6 Findings, recommendations and future work

6.1 Introduction

In this final chapter I firstly present summaries of my principal findings and novel contributions to knowledge. I then make recommendations for change in epidemiological and public health research data management policy and practice. Following these, I discuss the overall strengths and weaknesses of the Ph.D. project and explore potential future research directions.

6.2 Summary of Ph.D. findings

6.2.1 Research case study I: Enhancing data discoverability

Background: An increase in the linking together, and utilising of, disparate datasets is helping to maximise opportunities to investigate the origins of disease and influences on the life course. However, the limited discoverability of epidemiological and public health research data renders it challenging for researchers to identify these datasets as potential additional sources of data. This combined with an inability to sufficiently characterise these datasets due to limited provision of and access to metadata limits the extent to which these data be used.

Aim: To identify and evaluate mechanisms to enhance the discoverability of epidemiological and public health research data. The objectives were to: a) investigate current approaches to data discoverability; b) examine current stakeholders' awareness of issues relating to data discoverability and identify perceived challenges; and c) propose and evaluate methods to enhance data discoverability.

Methods: I used a combination of investigative techniques: a systematic literature review of current approaches to data discovery; an online stakeholder survey; and feasibility analyses of the mechanisms identified. I also used grounded theory to analyse the qualitative data to develop theories hermeneutically.

Results: A total of 49 public health and epidemiological studies and organisations were identified, 13 of which were reviewed. I identified varying approaches to facilitating data discovery; the most common format to present research protocols was Portable Document Format.

A total of 253 individuals completed the survey of which most undertake research in Europe. The survey identified the following perceived challenges with creating and/or using metadata: a) differing standards of research data management; and b) limited availability of resources. The survey also identified challenges associated with data publications such as limited academic incentives to produce these and a need for changes in research culture.

Based on findings from the literature review and survey, three models to enhance data discoverability were identified: a) data publications; b) semantic web technologies; and c) a public health portal. These were evaluated through the performance of feasibility studies and by engaging with stakeholders. The development of a portal proved most popular with stakeholders.

Conclusions: Data discovery is a fundamental step in the research data lifecycle and by implementing one or a combination of these mechanisms, researchers are more able to facilitate the discovery of their research data. The mechanism which proved most popular was the public health portal. Currently, there is no mandatory registration process for observational studies, this would be the first of its kind. The Wellcome Trust has begun preparations to take this work forward.

Furthermore, parts of the review, and online survey results and recommendations were published through the full and summary reports, (Castillo, Gregory et al. 2014). Elements of this work form part of a manuscript under review (McMahon, Denaxas et al. 2016). This work was also presented at the DDI conference NADDI in 2015.

6.2.2 Research objective II: Improving epidemiological and public health metadata quality assessment

Background: Making robust inferences from epidemiological and public health research data necessitates access to good quality and data lifecycle-based metadata. The provision of metadata can help researchers to better understand data and facilitate analyses; examples of metadata artefacts already being used by researchers in the epidemiological and public health domains include data dictionaries. However, in many instances, metadata are not subject to the same level of scrutiny as the research data they are associated with. Subsequently, metadata are of a variable quality and researchers cannot sufficiently characterise certain research datasets.

Aim: To create and evaluate a novel quality framework to assess metadata quality within epidemiological and public health research settings. The objectives were to: a) describe current practices in metadata quality assessment; b) identify metadata quality dimensions; c) create a novel quality assessment framework; and d) evaluate the framework through iterative application to test cases.

Methods: I used a combination of analytical techniques: a systematic literature review using cross-disciplinary databases to source literature on metadata quality assessment and an online stakeholder survey. The quality assessment framework was evaluated by applying it to three cohort studies as test cases and engaging with stakeholders.

Results: The performance of a systematic literature review combined with a comprehensive online stakeholder survey provided the evidentiary basis for the development of the novel metadata quality framework. The review identified 11 studies and nine dimensions of quality. The review did not identify a method of quality assessment designed for use in the epidemiology and public health domains.

A total of 96 individuals completed the survey globally most of whom were located in Europe. The survey identified challenges such as a lack of guidance to assist quality assessments associated with assessing metadata quality in epidemiology and public health. The survey also identified that ‘accuracy’ was deemed by respondents as the most important metadata quality dimension.

Results of both the systematic literature review and comprehensive online survey have shown that there is currently no framework designed specially to assess epidemiological and public health metadata quality. To address the challenges identified, I created a novel epidemiological and public health metadata quality assessment framework. The framework consists of four components: a) general information; b) tools and technologies; c) usability; and d) management and curation. The framework was evaluated by being applied iteratively to three test cases and engaging with stakeholders. The framework was evaluated by applying it iteratively to three test cases and by engaging with stakeholders.

Conclusions: Having access to good quality metadata can support researchers when sourcing additional datasets to investigate disease aetiology at scale. However, the literature review and survey demonstrated that metadata quality is variable and that there is currently no framework to assess metadata quality in the epidemiology and public health domains. By using the novel framework, researchers are now able to qualitatively assess epidemiological and public health metadata quality using a framework designed specifically for use in these domains. The framework has a unique basis in metadata quality dimensions such as discoverability and extendibility – dimensions not identified in any of the pre-existing frameworks. The framework also contains headings such as encoding and exchange standards (types of health information standard) again not identified in any of the existing frameworks. Use of this framework can also help to address issues relating to a lack of guidance through the explanations provided and can help researchers follow a formalised method of quality assessment.

Findings from this study were published in (McMahon, Castillo et al. 2015) and parts were also presented as a poster at a conference, (McMahon, Denaxas et al. 2015). Results of the online stakeholder survey and recommendations can be found in (McMahon and Denaxas 2016).

6.2.3 Research objective III: Improving recording of consent for record linkage metadata

Background: Researchers are increasingly linking disparate sources of data to produce datasets with an enhanced complexity and depth. These datasets may be harnessed to further investigate the longitudinal nature of disease; as such, they must be matched with consent models equally as enduring. Nevertheless, before these data may be linked together and utilised, researchers must harmonise these differing consent models. However, there is a lack of standardised methods to capture consent to record linkage metadata in longitudinal studies. By having standardised descriptions of this process, researchers can better understand the extent to which consent is given and under which conditions.

Aim: To create and evaluate a novel metadata management model to capture consent for record linkage metadata in longitudinal studies. The objectives were to: a) systematically identify and review methods for recording consent; b) comprehensively review consent elements from nine longitudinal studies; c) critically appraise DDI 3.2; and d) create and evaluate a novel metadata management model to record consent for record linkage metadata.

Methods: I performed a cross-disciplinary literature review to firstly source literature in this area. I then qualitatively analysed 30 consent forms from nine longitudinal studies to identify key consent elements. These elements were grouped together and a series of metadata management models were created using an object oriented modelling approach. I then critically appraised DDI 3.2 by mapping the key consent elements previously identified to those within DDI 3.2. These models were evaluated by being iteratively applied to three test cases.

Results: A total of 61 manuscripts were identified and reviewed all of which were categorised into themes inductively and iteratively. The literature review did not identify any manuscripts which described the development of metadata management models.

Following the analysis of the consent forms, the key consent elements identified were categorised into four groups: a) people; b) consent form; c) personal records; and d) information document. I created a corresponding metadata management model for each consent element group; all of which were applied to the test cases. The models were revised following each application.

Results of the critical appraisal of DDI 3.2 showed that whilst generalised metadata elements can be mapped directly, DDI 3.2 lacks the more specialist metadata elements needed to create detailed metadata descriptions of consent for record linkage.

I evaluated the model through its iterative application to a series of test cases. With each application, the model was tested and changes for improvement were made.

Conclusions: The novel metadata management model can now be used to assist stakeholders in documenting consent for record linkage given as part of a longitudinal study. This model also provides elements to assist researchers in creating standardised descriptions not found in the current prevailing standard, DDI 3.2.

Parts of this work were presented as posters at conferences, (McMahon, Dezateux et al. 2013a) and (McMahon, Dezateux et al. 2013b) and a publication is currently in press, (McMahon and Denaxas 2017).

6.3 Summary of novel contributions

This thesis makes multiple novel contributions to the field:

I. **Research case study I: Enhancing the discoverability of epidemiological and public health research data**

I.1. Performance of a systematic literature review and a large scale online stakeholder survey focusing on data discoverability; these investigations had not previously been undertaken in epidemiological and public health research settings.

I.2. Identified multiple areas of importance to data discovery, (3.4.2.7) such as identifying the commonalities between studies and being mindful of consent and other ethical issues pertaining specifically to the reuse of epidemiological and public health study data to maximise potential research opportunities.

I.3. Utilised a combination of techniques to identify and appraise three options, not currently applied to their fullest extent in epidemiological and public health research settings, to determine their strengths and weaknesses and make recommendations to better their application in these fields to enhance data discoverability.

II. **Research case study II: Improving metadata quality assessment in epidemiological and public health research**

II.1. Performance of a comprehensive literature review demonstrating that there had been no previous work undertaken on the metadata quality issue in epidemiological and public health research settings. The literature review took into consideration the wider implications of the metadata quality issue by evaluating metadata quality assessment criteria from other research areas and identifying metadata quality dimensions considered important by the wider research community.

II.2. Performance of an online stakeholder survey focusing on metadata and metadata quality assessment; an investigation not previously undertaken in epidemiological and public health research settings. The results of which informed the design and development of the novel metadata quality assessment framework.

II.3. Creation and evaluation of a novel metadata quality assessment framework for use in epidemiological and public health research settings. The framework addresses the gap in knowledge on assessing metadata quality in epidemiological and public health research settings.

III. Research case study III: Improving the recording of consent for record linkage metadata in longitudinal studies

III.1. Performance of a rigorous literature review demonstrating no work had previously been undertaken to address the challenges associated with recording consent for record linkage metadata in epidemiological and public health studies in a standardised manner by developing a standards-based model.

III.2. Critical appraisal of DDI3.2 evaluating the applicability and in particular the challenges of the current prevailing standard to record consent for record linkage metadata specifically. I fed the results of this appraisal into the design and development of the novel metadata management model by way of addressing the challenges associated with the recording of consent for record linkage metadata.

III.3. Creation and evaluation of a novel metadata management model which addresses the gap in knowledge on recording consent for record linkage metadata in a standardised and robust manner.

6.4 Recommendations

The findings from each research case study have informed a series of recommendations for change in epidemiological and public health research data management policy and practice. These have been categorised according to the stakeholder group at which they are aimed. The stakeholder groups are: a) all stakeholders; b) researchers, data users and data producers; c) archivists, data curators and librarians; and d) agencies – funding, governance, and formal academic review.

6.4.1 All stakeholders

I. Recommendation: Create a harmonised research data lifecycle model to support improved documentation of data and metadata work flows and to enhance stakeholders' understanding of the potential research benefits associated with a cyclical approach to research data management inclusive of lifecycle-based metadata

Background and problem: Researchers are increasingly recognising the potential research benefits associated with a cyclical approach to research data management. The current shift in research culture to adopt a cyclical approach to research data management must be matched with a cyclical approach to the way metadata are managed. Having a lifecycle-based approach to metadata markup will help to support research data reuse and repurposing across the stages of the research data lifecycle. However, given the heterogeneity of RDL models available and a lack of guidance as to when to use which one, systematic metadata markup is potentially hampered.

Basis in PhD research findings: My work specifically addresses this current lack of emphasis on a modular approach to the research data lifecycle and issues relating to metadata management within these individual modules. My discussion in 1.1.3 on a data lifecycle-based approach to epidemiological and public health studies provides the basis for the recommendation to create a harmonised set of RDL stages. This work builds on existing work focusing on the research data lifecycle; the novelty of my work lies in the particular focus placed on organising the different stages of the RDL into modules which can be addressed either individually, as demonstrated through my work in chapter 6 where I created and evaluated a novel metadata management model to standardise the recording of consent to record linkage in longitudinal studies, or a combination of stages as demonstrated by my work in chapter 5.

Recommendation: A harmonised set of RDL stages with definitions is needed so that metadata markup may become more modular in design and stakeholders are potentially more able to effectively compare metadata across the different stages. However, determining how granular definitions of each RDL stage could prove challenging. This is because data curation processes may differ according to the type of study being documented.

The development of a harmonised RDL model is recommended and stakeholders should identify at which stage their research study sits and begin generating metadata inclusive of data flows and any steps taken to manage the data. In having a clearly defined metadata pathway, this could help to ease the process of creating and maintaining metadata and help researchers to identify opportunities to produce data publications.

II. **Recommendation:** Improve awareness of the implications associated with poor quality metadata in public health and epidemiological research

Background and problem: High quality metadata is needed in epidemiological and public health research to not only describe research study results (Musen, Bean et al. 2015). In an article by Phimister (2015), the author discusses how missing metadata in genetic databases coupled with constraints over the sharing of metadata put in place due to the way in which consent was given, have resulted in incomplete descriptions of her family history and potentially incomplete descriptions of her genetic data possibility reducing scope for reuse. However, results of the survey (chapter 4.4.2.5) show one of the main challenges in assessing quality in public health and epidemiological research is a lack of awareness of the issue of poor quality metadata and the potential implications this can have on research and research data management.

Basis in PhD research findings: It is this problem of a limited awareness of poor quality metadata in public health and epidemiological research that my work on a novel metadata quality assessment framework addresses. Findings from chapter 4 and in particular sections, 4.4.2.2, 4.4.2.4, 4.4.2.5 provide evidentiary basis for this recommendation. For example, results of the survey chapter 4.4.2.2 showed that of those who submitted data, most regularly use metadata; yet, awareness of the metadata quality issue is limited (McMahon, Denaxas et al. 2016).

Recommendation: A change is needed in research culture to widen the focus to include the quality of metadata and to increase attention to the different dimensions of metadata such as completeness and accuracy. Researchers are dependent on research artefacts such as data dictionaries and so mechanisms are needed to better educate researchers on the potential negative outcomes of producing and/or using poor quality metadata in their research. Potential incentives to encourage researchers to adopt better metadata management practices include enhanced opportunities for their research data to be discovered and reused resulting in increased collaboration. A potential challenge with this is determining the how to quantifiably measure how effectively these mechanisms are working and their effect on the research environment.

III. Recommendation: Integrate metadata quality assessments into stakeholders' work routines as supported by increased provision of training and guidance

Background and problem: The provision of research data coupled with good quality metadata can enable stakeholders to glean even greater insights from datasets of varying sizes (Thornton 2015). For example, access to variable level metadata enabled Williams, Bunch et al. (2013) to develop a protocol maximising "…the sensitivity and specificity of the data linkage…" in their study looking at the risk of cancer in children following assisted conception. However, results of the survey show that currently, quality assessments are only performed sometimes (Table 4-12 and chapter 4.4.2.5) and in certain circumstances not at all.

Basis in PhD research findings: My work addresses the challenge of integrating quality assessment into stakeholders work routines by providing a tool – the novel framework – to help users assess metadata quality. The framework (chapter 4.3.3) created and evaluated can be adopted by stakeholders from across the epidemiological and public health research domains (McMahon, Denaxas et al. 2016).

Recommendation: The novel framework serves as a formalised approach to assist metadata quality assessment in the epidemiological and public health research domains. Inclusion of domain-specific detail not before seen in other pre-existing frameworks such as, 'use of clinical terminologies' and 'provision of links to other studies, sweeps or publications' are just two examples of how the framework has been tailored to meet the needs of epidemiological and public health metadata. There is scope to develop the framework further to include a set of quantitative metrics which complement the underlying metadata quality dimensions. By having a set of quantitative metrics, changes to the metadata may become easier to monitor and benchmark against other standardised metadata instances.

IV. **Recommendation:** Investigate mechanisms to further integrate health information standards into epidemiological and public health research

Background and problem: In chapter 2.3, I explored the realised research benefits of applying health information standards in epidemiological and public health research where they are not currently widely adopted. For example, stakeholders can use ontologies to map between ontological concepts for purposes of data discovery and data reuse (Subramanian, Kitchen et al. 2015). Furthermore, according to Dugan, Emrich et al. (2014), the provision of interoperable databases was enabled through use of a semantic framework which was ontology-based. This helped to maximise benefits associated with use of semantic frameworks and reduce potential information loss during transmission or conversion.

Basis in PhD research findings: My work in chapter 2 addressed this challenge by my investigating and discussing potential ways of utilising these standards so that the potential research benefits of integrating health information standards (encoding and exchange) may be harnessed. Results of the evaluation of the mechanisms in chapter 3.4.4 of the models have demonstrated that there are many potential research benefits associated with the application of information standards. Increased use of standards was also identified as an area of improvement and immediate priority in chapter 3.4.2.8. Other such benefits are discussed in Chapter 2.

Recommendation: By further investigating mechanisms to better integrate SWTs, potential benefits may be realised and stakeholders better supported in maximising the research potential of data. Nevertheless, the challenge here is providing adequate training and support to researchers instructing them how to effectively integrate these into their research practices. More resources are potentially needed here. This work builds on and furthers existing work on the improved application of standards in life sciences research. The novelty of my work lies in the focus on the application of health information standards, namely encoding and exchange standards, into epidemiological and public health research settings in support of a cyclical approach to research data management.

6.4.2 Researchers, data users, and data producers

V. Recommendation: Make better use data management plans as an approach for characterising research data to enhance the potential for data reuse and repurposing

Background and problem: Data management plans (DMPs) describe how research data will be managed over the course of a study. Good quality descriptions of research data are vital in helping stakeholders determine whether studies are feasible using a particular dataset without accessing the dataset(s) directly (Lee, Black et al. 2015). DMPs, if written well, can do just this. For example, DMPs can contain details relating to (potential) application of clinical terminologies; details of which are vital to developing cross-institutional solutions to problems associated with phenotyping (Shivade, Raghavan et al. 2014). DMPs are most commonly used when researchers are submitting grant applications to funding agencies. Nevertheless, use of DMPs to characterise research data beyond that which is expected of them during a grant review process is not commonplace and the plans themselves can be of variable quality; consequently limiting the extent of their potential use as a mechanism to enhance the characterisation of data.

Basis in PhD research findings: My work in chapter 5 on improving the way in which consent for record linkage metadata are recorded demonstrates how increased standardisation could improve its associated quality. If the way in which DMPs are produced were to be standardised, and these processes became a part of regular work routines, this could help to address the associated issue of variable quality.

Recommendation: It is recommended more emphasis is placed on improving the quality of DMPS and their integration into daily work routines of stakeholders. Stakeholders should be encouraged to consider the research data lifecycle and its impact on performing and documenting life sciences research. Development and management of DMPs should become an iterative process with each new version reflecting any changes made to the management strategy; organisations such as the Digital Curation Centre provide an online tool to help generate DMPs (DCC 2014).

This recommendation adds to current guidelines focusing on the use of DMPs by encouraging users to make greater use of existing resources i.e. a study's DMP. The novelty of my recommendation lies in the extension made tot eh scope of a DMP's various uses – now DMPs should be looked at as another potential tool which may be harnessed to characterise research data for reuse/repurposing.

VI. **Recommendation:** Increase identification of commonalities and links between studies through improved provision of openly available metadata and other associated research artefacts

Background and problem: Realising the benefits of collaborative working across institutions globally necessitates some form of data sharing. Data sharing within this context can be seen as a step towards making effective use of resources by helping to reduce duplicated efforts (Thiru, Hassey et al. 2003; Tenopir, Allard et al. 2011; Borgman 2012). However, certain research studies have involved researchers creating their own data elements list (Dugas, Jockel et al. 2015). Consequently, this is not always an effective use of resources when other, pre-existing lists may be repurposed and reused.

Basis in PhD research findings: My work addressees the current lack of focus on the impact a lack of openly available metadata could potentially have on the extent to which commonalities may be identified between studies. Survey results in chapter 3.4.2 provide the basis for the recommendation to increase the identification of commonalities and links between studies.

Recommendation: Researchers are advised to identify commonalities between studies as early on in the lifecycle as possible to help inform design and development of their own study protocols and to identify best practice (Castillo, Gregory et al. 2014). The development of a public health portal as discussed in chapter 3 would help to facilitate this as the portal would hold a collection of metadata records created through some kind of registration process. These records would ideally contain variable-level detail which can be used to help identify areas of potential overlap and reuse. Researchers should aim to record, and make openly available, metadata throughout the course of a study and to standardise where possible. However, much is still needed in the way of incentivising researchers to create detailed high quality metadata records which can then be shared. Researchers should also be more encouraged, again through incentives such as potential increase in citations through creation of data publications (as discussed as an area of improvement and priority in chapter 3.4.2.8), to publicise their research data to help maximise its reuse potential.

The recommendation adds to the research priorities as discussed by Moher, Glasziou et al. (2016) focusing on making, "publicly available the full protocols, analysis plans or sequence of analytical choices, and raw data for all designed and undertaken biomedical research." The novelty of my recommendation lies in users being encouraged to then utilise these openly available artefacts to then identify commonalities between studies with a view to repurposing research data to maximise their reuse opportunities.

6.4.3 Archivists and librarians

VII. Recommendation: Adopt better documented metadata pathways which map from where metadata have been sourced, how they have been used and where they have been shared inclusive of any additional research artefacts such as consent forms

Background and problem: Documenting research data without use of adequate and standardised methods to do so can hinder the extent to which these data are managed effectively (Parekh, Armananzas et al. 2015). Current unstandardised approaches to metadata markup can reduce the potential for data discovery and reuse (Toga, Foster et al. 2015).

Basis in PhD research findings: My work addresses the problems associated with a lack of standards-based approaches to manage metadata.

The basis for this recommendation can be found in chapter 5 and in particular sections, 5.1 and 5.2.1.

Recommendation: Mapping metadata pathways will help stakeholders to determine which tools/technologies work well, which need further development, and to facilitate the development of best practice guidelines within the context of metadata markup in epidemiological and public health research. More specifically, mapping the metadata pathway serves two purposes: a) helps to determine at which stages of the research data lifecycle the metadata, (plus its type and format) have been handled and which, if any, changes were made; and b) supports efforts to formalise an approach to integrating metadata quality assessments along the research data lifecycle. This recommendation adds to other recommendations made such as providing this kind of metadata will help improve stakeholders' understanding of the circumstances under which the metadata were created (Musen, Bean et al. 2015). Also, access to study documentation will also help to improve scope for data sharing without compromising participant confidentiality (Dugas 2013).

As CIMs define the structural and semantic details needed to produce a detailed description (Moreno-Conde, Moner et al. 2015), these could help the mapping process. Furthermore, lists of tools and technologies should be collated together to identify what has been used, for which purpose, and why these tools/technologies were selected. Tools and technologies can include use of clinical terminologies and other such controlled vocabularies, SWTs, and metamodels. In terms of sharing the metadata this includes stakeholders' manually sending metadata e.g. PDF through emails, and uploading and/or creating metadata records into public facing catalogues. It can also include use of any metadata harvesting protocols to expedite this process potentially reducing the risk of human error. Retrospective metadata markup is an alternative approach to markup in real time, but there is the potential increase in the risk of recall bias. Appropriate use of metadata markup tools could assist the process; survey results show in certain circumstances this is already happening.

6.4.4 Agencies - Funding, governance, and formal academic review

VIII. Recommendation: Improve recognition of the significance of data publications and other such published[5] articles in formal academic reviews

Background and problem: Data publications describe not only the research data but the ways in which the data were collected and managed. These papers could also include descriptions of study protocols and investigative/analytical methods. Currently, data publications are under-valued by the wider research community particularly within the context of formal academic reviews. And yet, these publications are critical to the enhanced discoverability of research data and support increased scope of data reuse and repurposing.

Basis in PhD research findings: My work in chapter 3 addresses the need to improve the perceived significance of data publications and other such articles by discussing their advantages to enhancing data discovery. The survey results, as detailed in chapter 3.4.2.6 demonstrate the need for increased academic rewards as incentives for the publishing of data publications and other such manuscripts.

Recommendation: More support is needed to move their position higher in the perceived hierarchy of academic literature. Increased focus is also needed on reducing the potential association stakeholders may make between the scientific value of a paper and its place of publication. This recommendation complements and adds to a recent recommendation calling for increased rewards including funding and recognition of stakeholders commitments to "…reproducible research and an efficient culture for replication of research" (Moher, Glasziou et al. 2016). A potential challenge here is determining how to encourage researchers to include data

[5] Published is defined here as available online and/or in print either through open access or through use of login credentials. The article must have a persistent unique identifier such as a DOI and be citable.

publications in submissions to formal academic review panels/bodies in addition to original research articles.

IX. Recommendation: Develop a public health portal to enable researchers to register observational studies for purposes of enhanced data discovery, repurposing and reuse

Background and problem: Greater opportunities are needed for the wider research community to make their research data and/or protocols more discoverable. However, to the best of my knowledge, there is no such registration process for observational studies. There are registration processes for other types of study such as clinical trials through ClinicalTrials.gov.

Basis in PhD research findings: My work addresses the current lack of some kind of formalised registration process for observational studies. A registration process such as this could help users to identify existing research studies by providing a centralised portal of metadata records. The basis for this recommendation can be found in the survey results (chapter 3.5.3) which demonstrate the need for a portal to enable stakeholders in public health and epidemiological research to register observational studies.

Recommendation: A potential way to incentivise stakeholders to do so is to increase use of metadata harvesting tools to automate this process. If there was some kind of pre-requisite or data use caveat that insisted on use of metadata catalogues to enhance the discoverability of the research data, this could potentially have a positive impact on knowledge of these catalogues and how they are used. The provision of meta-metadata will help support the management of metadata held within the catalogues. A challenge here is implementing the necessary infrastructure to enable these processes. Additional resources such as funding will also need to be arranged to facilitate this process. Furthermore, plans must be put in place to then maintain the public health portal and ensure longer term sustainability.

This portal will serve as a cross-funder initiative to enhance discoverability(Castillo, Gregory et al. 2014) and furthers a recent recommendation from Moher, Glasziou et al. (2016) stating “funders, sponsors, regulators… should endorse and enforce study registration policies”. It is this registration process that my work addresses.

6.5 Ph.D. thesis strengths and weaknesses

6.5.1 Literature reviews

Strengths: The strength of the literature reviews lie in how the literature was systematically sourced using both biomedical and computer science databases, sources of gray literature such as search engines and forward citation tracking (Kuper, Nicholson et al. 2006). For research case study 1, the use of the six point agreed criteria enabled me to thoroughly investigate current approaches to discoverability and better my understanding of the current challenges associated with data discovery. For the second and third research case studies, use of the PRISMA checklist (Moher, Liberati et al. 2009) enabled me to identify and review manuscripts in a systematic manner. Additionally, by inductively and iteratively identifying the themes, this helped to ensure a thorough and comprehensive analysis of the literature review findings.

Weaknesses: The weakness of the literature reviews lie in the restrictions put in place by the inclusion criteria such as needing to be available in English. Literature was also excluded if the publication was not openly accessible or available by logging into the journal using institutional login details. This could have excluded potentially useful literature for inclusion in the review. Other limitations include, using search terms specific to public health and epidemiology whilst the databases utilised were not limited to these two domains – they included other domains such as computer science, librarianship and archiving. Therefore, possible differences in use of controlled vocabularies may have resulted in literature being missed. Nevertheless, in adopting a systematic approach to the review process, the reviews were conducted in robust manner using an internationally recognised methodology.

6.5.2 Online data collection – stakeholder surveys

Strengths: The strength of the method of data collection for the first and second research case studies lie in my being able to request participation from stakeholders both within the UK and internationally in a relatively short period of time. The use of mailing lists enabled me to contact stakeholders in public health and epidemiological research quickly, and the request to forward the invitational email served to potentially increase the number of potential participant reached. Furthermore, in having web-based data collection, the potential risk of participants being unable to submit data in person due to their distance from London is reduced.(Schleyer and Forrest 2000; Whitehead 2011) Consequently, I received 96 responses to the metadata quality survey and 253 responses to the data discoverability survey. Additionally, in using a web-based survey, built-in mechanisms for data validation helped to reduce the potential risk of nonsensical information being recorded. The use of self-administered surveys helped me to maintain a standardised approach to data capture as use of interviewers, each with potentially varying interviewing techniques was removed (Coggon, Rose et al. 2014). Having in place mechanisms to involve international stakeholders adds to the potential applicability of my findings on a global scale as my results and recommendations are based on both a national and international perspective.

Additionally including open-ended questions in the surveys helped me to capture qualitative data (MacKenzie, Wyatt et al. 2012). According to a study by Keusch (2014) on the design of self-administered, web-based questionnaires, the length of the input field of an open-ended question can impact how detailed the response provided is. Hence, to guide the participants when submitting their data through the surveys, one of the five potential ways to structure input fields to open-ended questions in web surveys as suggested by Couper, Kennedy et al. (2011) are included. The style of question most suitable to the surveys was a narrative response with no length or formatting constraints. This is so the participants have the freedom to provide additional answers, describe their experiences, and share their opinions without restriction.

Weaknesses: Having used an online data collection method, a potential weakness here is the possibility that my data is biased towards those with online access. The use of telephone assisted or postal submissions may have led to different outcomes. Furthermore, by using mailing lists, I was dependent on those managing the lists to have current and working emails addresses for all those that subscribe, and that those subscribed to these lists read their email messages. Also, by requesting those that receive the email to forward it to other potentially interested parties, determining a response rate is unfeasible.

In addition, there is also the possibility for one individual to have submitted multiple answer sets as the only unique identifier assigned to a set of answers is a record ID once the responses have been received by REDCap - the survey system does not monitor who has already submitted data. Though this is a potential weakness in my approach to data collection, in not recording any kind of identifiable information from the respondent, the risk of unintentionally identifying the respondent through the responses provided is potentially reduced.

6.5.3 Qualitative data analyses – a Grounded Theory approach

Strengths: When analysing the qualitative results for data discoverability and metadata quality studies collected through the online stakeholder surveys, I used a Grounded Theory approach (Glaser and Strauss 1967). The strength in using Grounded Theory to analyse qualitative results lies in being able to collect rich data (Glaser 1999) more reflective of reality. This theory focuses on studying the interactions of the respondents when exposed to a particular situation (Sbaraini, Carter et al. 2011). Grounded Theory enabled me to explore these data without having to adhere to pre-defined theories; rather, I was able to develop my own ideas (Callen, Braithwaite et al. 2008; Thomassen, Ahaus et al. 2014). The freedoms that came with applying a Grounded Theory approach enabled me to gather data, and then analyse these data without having to determine whether I am proving or disproving pre-formulated hypotheses.

Qualitative research in biomedicine is increasing(Kuper, Reeves et al. 2008) and in using this approach I was able to perform my data analyses guided by an established approach applied regularly across healthcare research (Mays and Pope 2000; Foley and Timonen 2015). When analysing my data, I adopted a hermeneutic approach and identified and grouped themes inductively and iteratively. In adopting this approach, I was able to build theories in a systematic and robust manner and was able to identify subgroups within categories.

In terms of how well the results may be replicated, the process of resurveying stakeholders online is fairly straightforward; a potential barrier is gathering respondents with the same characteristics and prior experience. Whilst I acknowledge that replicating results in qualitative research is a challenging process, in adopting a methodical approach to data collection, I was able to successfully gather data and develop theories. These theories in turn helped me to better my understanding of, for example, perceived barriers to enhancing data discoverability. It is this process of iteratively collecting ideas and themes which may be replicated by other researchers if the data are collected in the same way and of the same quality (Collingridge and Gantt 2008).

Weaknesses: There are several limitations to applying a Grounded Theory approach. As the basis of these qualitative analyses is interpretive, this could reduce the generalisability (Winkelman, Leonard et al. 2005) and applicability of research findings to the wider scientific community. Therefore, there is the possibility of potentially reduced external validity of my research findings. The generalisability of findings which are the result of using a Grounded Theory approach is improved through increased abstraction of findings (Corbin and Strauss 1990). It is possible, that my findings may need to become more abstract to increase their applicability to the wider scientific research community. Verifying research findings is a step towards reducing subjectivity and so more is potentially needed to increase objectivity in my findings (Corbin and Strauss 1990). Nevertheless, in following a robust method of data analysis, I was able to successfully derive theories from the data collected and used these to inform my recommendations for change in policy and practice.

6.5.4 Models/framework design and development

Strengths: The strength in using a standards-based approach to designing and developing the models and framework lies in the ability to harness existing standards and their scalability to build formalised models and frameworks for application in epidemiological and public health research settings. These standards can be critically appraised, as demonstrated in chapter 5.6.3, and extended to better enable their application in epidemiological and public health research. For example, an object oriented approach was used to develop the metadata management models to record consent metadata. Consequently, simple yet detailed models were quickly created which can be further developed. Further, an adherence to object oriented principles such as inheritance enabled me to reduce the risk of the model including repeating or redundant elements. This is advantageous as the model created and evaluated is more efficient and the scope for it to be instantiated is potentially increased.

Weaknesses: A potential weakness in the design and development of the quality assessment framework lies in the lack of quantitative quality analysis of the metadata. Currently, stakeholders are expected to perform qualitative analyses using the framework. Though this will work in the short term, the longer term goal is to produce a series of computable metrics to automate the assessment process. These algorithm-based mechanisms aim to quantify the assessment process helping to improve its objectivity by reducing any potential bias in the findings and quicken the overall assessment process. The metadata quality dimensions these metrics could prove potentially useful in assessing more effectively include completeness and accuracy. Furthermore, quantitative findings (as opposed to the current qualitative findings), can be better compared. By making these comparisons, areas of potential need of further support can be provided. Nevertheless, this current lack of quantitative metrics is demonstrative of the challenges associated with providing low level definitions of metadata quality dimensions which would form the basis of quantitative analysis.

6.5.5 Evaluation strategies

Strengths: the strength of the evaluation strategy for the first research case study lies in performing a series of feasibility analyses following an established methodology. Feasibility analyses originate from the computer science domain and promote a systematic evaluation of the three key aspects of the mechanisms - technical, organisational, and economic feasibility within the context of enhancing the discoverability of observational studies. For the second and third research case studies, the metadata quality assessment framework and consent for record linkage metadata management model, were applied to a series of test cases and improved after each iteration. These real world applications enabled a thorough testing of the quality framework and metadata management model.

Weaknesses: The weakness in conducting feasibility analyses for the first research case study lies in them not generally being applied in epidemiological and public health research settings. Feasibility analyses are generally associated with object oriented systems analysis and designs as a decision support tool when determining whether to buy and/or develop new tools. As this method was not specifically designed for use in epidemiological and public health research context, it is possible that aspects of each mechanism may not have been as fully explored as possible.

The weakness in the evaluation strategy for the third research case study lies in the lack of stakeholder engagement. The models were iteratively applied to test cases only and improved after each application. Nevertheless, this approach to evaluation is demonstrative of the advantage of using real world applications of model as these can be shared with stakeholders to facilitate debate and discussion around the models and their application in epidemiological and public health research.

6.6 Future direction

Research case study I: The next step is to begin planning the development of the public health portal. I will work with the community to determine the requirements of the portal and how stakeholders would ideally interact with this tool. Tasks I am likely to undertake include: a) creating and evaluating conceptual models of the system to circulate to stakeholders; b) creating and evaluating an underlying metadata model to which the metadata records will comply; c) identifying and critically apprising metadata standards to incorporate into the public health portal.

In addition to building the portal, I will continue to investigate the application of semantic web technologies and data publications to epidemiological and public health research. This is so I can harness their advantages and feed them into the portal where possible. For example, I will link to any associated data publications for each study from their metadata record. I will also further investigate the challenges associated with using semantic web technologies and data publications so that I can continue to address these. I will then be able to provide a set of recommendations to facilitate their enhanced application in epidemiological and public health research settings.

Research case study II: The next step for the metadata quality study is to investigate a set of quantitative measures of quality. These measures will help to increase objectivity by reducing the amount of assessment conducted manually. To implement these quantitative metrics, I will firstly determine if there are any pre-existing quantitative metadata metrics for use in epidemiological and public health research settings. If any are in existence, I will critically appraise these for use looking particularly at where extensions are needed before they may be applied to the framework. If no such metrics currently available, I will investigate computational techniques to determine how algorithm-based mechanisms may be harnessed to better determine metadata quality. In applying these mechanisms to the framework, I will also be able to quicken the overall assessment process to potentially increasing efficiency.

Once I have conducted the systematic literature review, critically appraised (if any) existing metrics, and developed novel metrics where needed, I will then extend the scope of the framework by applying it to metadata associated with cohorts identified using electronic healthcare records. It is possible that the metadata associated with data whose primary use is research, and data whose primary use is to assist clinical care, perform differently in terms of quality according to the framework. For example, when assessing use of encoding standards such as ICD 10, stakeholders are dependent on clinical information having been encoded accurately and in a timely manner. An example of where clinical information have been encoded using a standard is HES data. However, it is possible that potential differences in the approach taken to encode this information between that collected from existing clinical data and that from research studies could impact their associated metadata and its quality. A key element here is deciphering whether the metadata quality varies due to the process of metadata maintenance and curation, or whether, the metadata of repurposed clinical data (resulting from clinically phenotyped cohorts) cannot be assessed by the same framework. Further investigation is needed to determine the cause of these potential differences and any subsequent implications particularly on the use of quantitative metrics as these may need to be altered too.

Research case study III: The next step for the consent for record linkage metadata study is to design a user interface to assist stakeholders in documenting their metadata. This will help stakeholders to interact with the model and ease its application into daily work routines. The first step would be to hold a series of focus groups to present the model and gain feedback particularly around areas in need of further development. The information and critiques gathered from these focus groups will also feed into the user requirements I will need to collect to ensure my user interface is fit for use.

The tool would be built on the model and incorporate the corresponding XML schema. I will also investigate ways to tailor the user experience when using the tool. For example, if the user indicates that there will be no face-to-face interviews, any additional metadata elements relating to face-to-face interviews will be hidden from view. In doing so, the user will only be presented with the metadata fields relevant to their study. Nevertheless, by ensuring that the metadata fits the underlying model, future versions of the metadata, where additional metadata elements are needed, can be included without previous versions being negatively impacted such as introduction of missing or distorted metadata.

6.7 Closing remarks

As a biomedical informatician, I am neither computer scientist nor clinician; I sit in the space between. My research focuses on creating solutions to informational workflow problems in epidemiological and public health research settings. In this body of work, I demonstrated how metadata management models may be applied to enhance the use and reuse of research data and made multiple novel contributions to knowledge. The recommendations made in this thesis aim to enhance the research environment by addressing current challenges in epidemiological and public health research data management policy and practice.

References

. "HITECH Act." Retrieved February 18, 2014, from http://www.healthit.gov/policy-researchers-implementers/hitech-act.

(1996). Health Insurance Portability and Accountability Act of 1996. United States of America.

(2009). Metadata and Semantics. New York, Springer Science+Business Media, LLC.

(2013a). Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data.

(2013b). Users' Guide. C. McDonald, S. Huff, J. Deckard, K. Holck and D. J. Vreeman.

(2014a). "BioPortal." Retrieved February 25, 2014, from http://bioportal.bioontology.org/.

(2014b). "BioSharing." Retrieved February 24, 2014, from http://biosharing.org/.

(2014c). "DSM: History of the Manual." Retrieved May 07, 2014.

(2014d). "LOINC Overview." Retrieved May 07, 2014, from http://loinc.org/faq/getting-started/getting-started/#how-are-loinc-and-relma-distributed.

(2014e, 2014 March 11). "MeSH Browser." Retrieved May 07, 2014, from https://www.nlm.nih.gov/mesh/MBrowser.html.

(2014f). "RDF 1.1 Concepts and Abstract Syntax." Retrieved September 22, 2015, from http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

(2014g, 2014 April 29). "SNOMED CT Release Files." Retrieved May 07, 2014, from http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html.

(2014h). "UK Data Service." Retrieved June 24, 2014, from http://ukdataservice.ac.uk/.

(2014i, 2014 January 16). "Unified Modeling Language (UML)." Retrieved February 07, 2014, from http://www.uml.org/.

(2015a). "Clinical Information Modeling Initiative (CIMI)." Retrieved October 09, 2015, from http://www.opencimi.org/.

(2015b). "Health Level Severn International." Retrieved September 22, 2015, from http://www.hl7.org/.

(2015c). "International Classification of Diseases (ICD)." Retrieved September 22, 2015, from http://www.who.int/classifications/icd/en/.

(2015d). "Protégé." Retrieved October 7, 2015, from http://protege.stanford.edu/products.php.

(2016a). "Alzheimer's Disease Neuroimaging Initiative." Retrieved July 19, 2016, from http://www.adni-info.org/.

(2016b). "Avon Longitudinal Study of Parents and Children." Retrieved February 4, 2016, from http://www.bristol.ac.uk/alspac/.

(2016c). "Born in Bradford." Retrieved February 4, 2016, from http://www.borninbradford.nhs.uk/.

(2016d). "British Household Panel Survey." Retrieved February 4, 2016 from https://www.iser.essex.ac.uk/bhps.

(2016e). "CESSDA." Retrieved February 09, 2016, from http://cessda.net/.

(2016f). "Clinical Practice Research Datalink." Retrieved July 19, 2016, from https://www.cprd.com/intro.asp.

(2016g). "CLOSER." Retrieved February 09, 2016, from http://www.closer.ac.uk/.

(2016h). "Emerging Risk Factors Collaboration ". Retrieved July 19, 2016, from http://www.phpc.cam.ac.uk/ceu/research/erfc/.

(2016i). "Hospital Episode Statistics." Retrieved July 25, 2016, from http://www.hscic.gov.uk/hes.

(2016j). "IHSN Survey Catalog." Retrieved February 09, 2016, from http://catalog.ihsn.org/index.php/catalog.

(2016k). "Journal of Open Health Data." Retrieved July 25, 2016, from http://openhealthdata.metajnl.com/.

(2016l). "Life Study." Retrieved February 4, 2016, from http://www.lifestudy.ac.uk/homepage.

(2016m). "Open Journal of Bioresources." Retrieved July 25, 2016, from http://openbioresources.metajnl.com/.

(2016n). "PLOS." Retrieved July 19, 2016, from https://www.plos.org/.

(2016o). "Research Excellence Framework." Retrieved July 19, 2016, from http://www.ref.ac.uk/.

(2016p). "UK Biobank." Retrieved February 4, 2016, from http://www.ukbiobank.ac.uk/.

(2016q). "Understanding Society." Retrieved February 4, 2016, from https://www.understandingsociety.ac.uk/.

Administrative Data Taskforce (2012). The UK Administrative Data Research Network: Improving Access for Research and Policy. UK.

Akkad, A., C. Jackson, et al. (2006). "Patients' perceptions of written consent: questionnaire study." BMJ **333**(7567): 528.

Al-Enazi, T. and S. El-Masri (2013). "HL7 Engine Module for Healthcare Information Systems." J Med Syst **37**(6): 9986.

Al-Shorbaji, N. (2012). WHO Forum on Health Data Standardization and Interoperability. WHO Forum on Health Data Standardization and Interoperability. Switzerland.

American Psychiatric Association (2014). "DSM-5 Development." Retrieved January 30, 2014, from http://www.dsm5.org/Pages/Default.aspx.

Anderson, N. R., E. S. Lee, et al. (2007). "Issues in biomedical research data management and analysis: needs and barriers." J Am Med Inform Assoc **14**(4): 478-488.

Anwar, N. and E. Hunt (2009). "Francisella tularensis novicida proteomic and transcriptomic data integration and annotation based on semantic web technologies." BMC Bioinformatics **10 Suppl 10**: S3.

Armstrong, D., E. Kline-Rogers, et al. (2005). "Potential impact of the HIPAA privacy rule on data collection in a registry of patients with acute coronary syndrome." Arch Intern Med **165**(10): 1125-1129.

Arora, A., S. Rajagopalan, et al. (2011). "Development of tool for the assessment of comprehension of informed consent form in healthy volunteers participating in first-in-human studies." Contemp Clin Trials **32**(6): 814-817.

Australian Insitute of Health and Welfare (2014). "About METeOR." Retrieved July 18, 2014, from http://meteor.aihw.gov.au/content/index.phtml/itemId/181414.

Ball, A. and M. Duke (2012). Data Citation and Linking. Edinburgh.

Belleau, F., M.-A. Nolin, et al. (2008). "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems." Journal of Biomedical Informatics **41**(5): 706-716.

Berry, J. G., P. Ryan, et al. (2011). "A randomised controlled trial to compare opt-in and opt-out parental consent for childhood vaccine safety surveillance using data linkage: study protocol." Trials **12**: 1.

Berry, J. G., P. Ryan, et al. (2012). "A randomised controlled trial to compare opt-in and opt-out parental consent for childhood vaccine safety surveillance using data linkage." J Med Ethics **38**(10): 619-625.

Bian, J., M. Xie, et al. (2014). "CLARA: an integrated clinical research administration system." Journal of the American Medical Informatics Association.

BioSharing (2014a). "Databases." Retrieved February 24, 2014, from http://biosharing.org/biodbcore.

BioSharing (2014b). "Policies." Retrieved February 24, 2014, from http://biosharing.org/policies.

BioSharing (2014c). "Standards." Retrieved February 24, 2014, from http://biosharing.org/standards.

Bloomrosen, M. and D. E. Detmer (2010). "Informatics, evidence-based care, and research; implications for national policy: a report of an American Medical Informatics Association health policy conference." J Am Med Inform Assoc **17**(2): 115-123.

Blumenthal, D. (2010). "Launching HITECH." New England Journal of Medicine **362**(5): 382-385.

Blumenthal, D. (2011). "Wiring the Health System — Origins and Provisions of a New Federal Program." New England Journal of Medicine **365**(24): 2323-2329.

Borgman, C. L. (2012). "The conundrum of sharing research data." Journal of the American Society for Information Science and Technology **63**(6): 1059-1078.

Boulton, G., P. Campbell, et al. (2012). Science as an open enterprise. London, The Royal Society

Brazma, A., M. Krestyaninova, et al. (2006). "Standards for systems biology." Nat Rev Genet **7**(8): 593-605.

BRIDG (2012a, 2012 September 5). "BRIDG." Retrieved November 20, 2013, from http://www.bridgmodel.org/.

BRIDG (2012b). "How is BRIDG updated? What is harmonization?". Retrieved November 20, 2013, from http://bridgmodel.nci.nih.gov/faq/browse-faqs-1/harmonization_of_BRIDG_Model/.

BRIDG (2012c, 2012 June 26). "Who is using BRIDG? How?". Retrieved January 13, 2014, from http://bridgmodel.nci.nih.gov/faq/browse-faqs-1/using_BRIDG_Model/.

British Standards Institution (2007). BS ISO/HL7 21731: 2006: Health Informatics - HL7 version 3 - Reference information model - Release 1 London, British Standards Institution.

Bruce, T. and D. Hillmann (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. Metadata in Practice. Chicago, American Library Association**:** 238-256.

BS EN ISO (2012). Health informatics - Electronic health record communication - Part 1: Reference model (ISO 13606-1:2008), BSI.

Buckles, V. D., K. K. Powlishta, et al. (2003). "Understanding of informed consent by demented individuals." Neurology **61**(12): 1662-1666.

Buckow, K., M. Quade, et al. (2014). "Changing Requirements and Resulting Needs for IT-Infrastructure for Longitudinal Research in the Neurosciences." Neurosci Res.

Budin-Ljosne, I., A. Tasse, et al. (2011). "Bridging consent: from toll bridges to lift bridges?" BMC Med Genomics **4**.

Bui, Y. and J.-r. Park (2006). "An assessment of metadata quality: A case study of the national science digital library metadata repository."

Burns, K. E. A., N. M. Magyarody, et al. (2011). "Attitudes of the general public toward alternative consent models." American Journal of Critical Care **20**(1): 75-83.

Callen, J. L., J. Braithwaite, et al. (2008). "Contextual implementation model: a framework for assisting clinical information system implementations." J Am Med Inform Assoc **15**(2): 255-262.

Castillo, T., A. Gregory, et al. (2014). Enhancing Discoverability of Public Health and Epidemiology Research Data. London, United Kingdom.

Caulfield, T., R. E. G. Upshur, et al. (2003). "DNA databanks and consent: a suggested policy option involving an authorization model." BMC Medical Ethics **4**(1): 1.

CDISC (2013a). "BRIDG Release 3.2 Now Available." Retrieved November 29, 2013, from http://www.cdisc.org/stuff/contentmgr/files/0/f900128a866087234d36b000198db840/misc/bridg_3_2_announcement_12___september__2012_final.pdf.

CDISC (2013b). "Mission & Principles." Retrieved November 20, 2013, from http://www.cdisc.org/mission-and-principles.

CDISC (2013c). "Operational Data Model." Retrieved November 29, 2013, from http://www.cdisc.org/odm.

Chulada, P. C., H. L. Vahdat, et al. (2008). "The Environmental Polymorphisms Registry: a DNA resource to study genetic susceptibility loci." Hum Genet **123**(2): 207-214.

Chute, C. G., M. Ullman-Cullere, et al. (2013). "Some experiences and opportunities for big data in translational research." Genet Med.

Ciccarese, P., E. Wu, et al. (2008). "The SWAN biomedical discourse ontology." Journal of Biomedical Informatics **41**(5): 739-751.

Cimino, J. J. (2011). "High-quality, standard, controlled healthcare terminologies come of age." Methods Inf Med **50**(2): 101-104.

Coggon, D., G. Rose, et al. (2014). Planning and conducting a survey. Epidemiology for the uninitiated, BMJ.

Coiera, E. and R. Clarke (2004). "e-Consent: The Design and Implementation of Consumer Consent Mechanisms in an Electronic Environment." Journal of the American Medical Informatics Association **11**(2): 129-140.

Collingridge, D. S. and E. E. Gantt (2008). "The quality of qualitative research." Am J Med Qual **23**(5): 389-395.

Connelly, R. and L. Platt (2014). "Cohort Profile: UK Millennium Cohort Study (MCS)." Int J Epidemiol **43**(6): 1719-1725.

Corbin, J. and A. Strauss (1990). "Grounded Theory Research: Procedures, Canons and Evaluative Criteria." Zeitschrift für Soziologie **19**(6): 418-427.

Couper, M. P., C. Kennedy, et al. (2011). "Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys." J Off Stat **27**(1): 65-85.

Coveney, P. V. (2005). "Scientific Grid computing." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **363**(1833): 1707-1713.

Da Rocha, A. C. and J. A. Seoane (2008). "Alternative consent models for biobanks: The new Spanish law on biomedical research." Bioethics **22**(8): 440-447.

Dahlberg, K. (2010). "The essence of essences / the search for meaning structures inphenomenological analysis of lifeworld phenomena." 2010 **1**(1): 9.

Daniel, C., A. Sinaci, et al. (2014). "Standard-based EHR-enabled applications for clinical research and patient safety: CDISC - IHE QRPH - EHR4CR & SALUS collaboration." AMIA Jt Summits Transl Sci Proc **2014**: 19-25.

Davies, J., J. Gibbons, et al. (2014). "The CancerGrid experience: Metadata-based model-driven engineering for clinical trials." Science of Computer Programming **89, Part B**(0): 126-143.

Davis, N., A. Pohlman, et al. (2003). "Improving the process of informed consent in the critically ill." JAMA **289**(15): 1963-1968.

DCC (2014). "DMP Online ". Retrieved June 24, 2014, from https://dmponline.dcc.ac.uk/.

DCMI (2015). "Dublin Core Metadata Initiative." Retrieved September 17, 2015, from http://dublincore.org/.

DDI Alliance (2015a). "DDI-Codebook." Retrieved September 22, 2015, from http://www.ddialliance.org/Specification/DDI-Codebook/.

DDI Alliance (2015b). "DDI Lifecycle." Retrieved September 22, 2015, from http://www.ddialliance.org/Specification/DDI-Lifecycle/.

de Carvalho, E. C., A. P. Batilana, et al. (2010). "Application description and policy model in collaborative environment for sharing of information on epidemiological and clinical research data sets." PLoS ONE **5**(2): e9314.

de Vries, J., T. Williams, et al. (2014). "Knowing who to trust: exploring the role of 'ethical metadata' in mediating risk of harm in collaborative genomics research in Africa." BMC Med Ethics. 2014 Aug 13; 15:62

Denaxas, S. C., J. George, et al. (2012). "Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)." Int J Epidemiol **41**(6): 1625-1638.

Dennis, A., B. H. Wixom, et al. (2015). Systems Analysis and Design. USA, John Wiley & Sons.

Deus, H. F., E. Prud'hommeaux, et al. (2012). "Translating standards into practice – One Semantic Web API for Gene Expression." Journal of Biomedical Informatics **45**(4): 782-794.

Digital Curation Centre (2014). "Biology." Retrieved September 25, 2014, from http://www.dcc.ac.uk/resources/subject-areas/biology.

Donnan, P., D. McLernon, et al. (2009). "Development of a decision support tool for primary care management of patients with abnormal liver function tests without clinically apparent liver disease: a record-linkage population cohort study and decision analysis (ALFIE)." Health Technology Assessment **13**(25).

Dugan, V. G., S. J. Emrich, et al. (2014). "Standardized metadata for human pathogen/vector genomic sequences." PLoS ONE **9**(6): e99979.

Dugas, M. (2013). "Re: Sharing data from clinical trials: where we are and what lies ahead." BMJ **347:f4794**.

Dugas, M., K. H. Jockel, et al. (2015). "Memorandum "Open Metadata". Open Access to Documentation Forms and Item Catalogs in Healthcare." Methods Inf Med **54**(4): 376-378.

Dumontier, M., AJG. Gray, et al. “The health care and life sciences community profile for dataset descriptions.” PeerJ 2016 Aug 16;4:e2331 https://doi.org/10.7717/peerj.2331

Dunn, K. M., K. Jordan, et al. (2004). "Patterns of consent in epidemiologic research: Evidence from over 25,000 responders." Am J Epidemiol **159**(11): 1087-1094.

Economic & Social Research Council (2013, 2013 June 14). "ESRC welcomes Government response to Administrative Data Taskforce report." Retrieved September 22, 2013, from http://www.esrc.ac.uk/news-and-events/announcements/26572/esrc-welcomes-government-response-to-administrative-data-taskforce-report.aspx.

Elger, B. S., J. Iavindrasana, et al. (2010). "Strategies for health data exchange for secondary, cross-institutional clinical research." Computer Methods and Programs in Biomedicine **99**(3): 230-251.

Elings, M. W. and G. Waibel (2007). Metadata for all: Descriptive standards and metadata sharing across libraries, archives and museums, Munksgaard International Publishers.

Ellul C, Foord J, Mooney J. Making metadata usable in a multi-national research setting. Appl Ergon. 2013 Nov;44(6):909-18.

Embi, P. J., C. Weir, et al. (2013). "Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators." Journal of the American Medical Informatics Association **20**(4): 718-726.

European Commission (2012, 2012 July 17). "Commission Recommendation of 17.7.2012 on access to and preservation of scientific information." Retrieved July 11, 2014, from http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf.

European Commission (2013). Study Report. eHealth - European Interoperability Framework. Luxembourg, Publications Office of the European Union

Eysenbach, G. (2004). "Improving the quality of Web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES)." J Med Internet Res **6**(3): e34.

Fernandez-Breis, J. T., J. A. Maldonado, et al. (2013). "Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts." J Am Med Inform Assoc.

Fernando, B., R. Bhojwani, et al. (2007). "Standards in consent for cataract surgery." J Cataract Refract Surg **33**(8): 1464-1468.

Foley, G. and V. Timonen (2015). "Using Grounded Theory Method to Capture and Analyze Health Care Experiences." Health Serv Res **50**(4): 1195-1210.

Freimuth, R. R., E. T. Freund, et al. (2012). "Life sciences domain analysis model." Journal of the American Medical Informatics Association **19**(6): 1095-1102.

Friedlander, J. A., G. S. Loeben, et al. (2011). "A novel method to enhance informed consent: a prospective and randomised trial of form-based versus electronic assisted informed consent in paediatric endoscopy." J Med Ethics **37**(4): 194-200.

Gefenas, E., V. Dranseika, et al. (2012). "Turning residual human biological materials into research collections: Playing with consent." J Med Ethics **38**(6): 351-355.

Glaser, B. and A. Strauss (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine.

Glaser, B. G. (1999). "The Future of Grounded Theory." Qual Health Res **9**(6): 836-845.

Gold, R. L., R. R. Lebel, et al. (1993). "Model consent forms for DNA linkage analysis and storage." Am J Med Genet **47**(8): 1223-1224.

Gonçalves, M. A., B. L. Moreira, et al. (2007). "“What is a good digital library?” – A quality model for digital libraries." Information Processing & Management **43**(5): 1416-1437.

Gori, S., M. T. Greco, et al. (2012). "A new informed consent form model for cancer patients: preliminary results of a prospective study by the Italian Association of Medical Oncology (AIOM)." Patient Educ Couns **87**(2): 243-249.

Gracie, S. K., A. W. Lyon, et al. (2010). "All Our Babies Cohort Study: Recruitment of a cohort to predict women at risk of preterm birth through the examination of gene expression profiles and the environment." BMC Pregnancy and Childbirth **10**.

Gray, L., G. D. Batty, et al. (2010). "Cohort Profile: The Scottish Health Surveys Cohort: linkage of study participants to routinely collected records for mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies." Int J Epidemiol **39**(2): 345-350.

Greenberg, J., M. C. Pattuelli, et al. (2006). "Author-generated Dublin Core metadata for web resources: a baseline study in an organization." Journal of Digital Information **2**(2).

Gregory, A. and W. Thomas (2012). From The Bottom Up.

Hammerschmidt, D. E. and M. A. Keane (1992). "Institutional Review Board (IRB) review lacks impact on the readability of consent forms for research." Am J Med Sci **304**(6): 348-351.

Harris, P. A., R. Taylor, et al. (2009). "Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support." Journal of Biomedical Informatics **42**(2): 377-381.

Haux, R., P. Knaup, et al. (2007). "On educating about medical data management: The other side of the electronic health record." Methods of Information in Medicine **46**(1): 74-79.

Health Level Seven International (2013). "HL7 Version 3: Reference Information Model (RIM)." Retrieved November 12, 2013, from http://www.hl7.org/implement/standards/product_brief.cfm?product_id=77.

Heidorn, P. B. (2008). "Shedding Light on the Dark Data in the Long Tail of Science." Library Trends **57**(2): 280-298.

Henderson, G. E. P. (2011). "Is Informed Consent Broken? [Article]." American Journal of the Medical Sciences October **342**(4): 267-272.

Hens, K., C. E. Van El, et al. (2013). "Developing a policy for paediatric biobanks: principles for good practice." Eur J Hum Genet **21**(1): 2-7.

Herrett, E., L. Smeeth, et al. (2010). "The Myocardial Ischaemia National Audit Project (MINAP)." Heart **96**(16): 1264-1267.

HL7 (2013). "Introducing HL7 FHIR." Retrieved January 16, 2014, from http://www.hl7.org/implement/standards/fhir/summary.html.

HL7 (2014, 2013 April 26). "Clinical Document Architecture (CDA)." Retrieved January 17, 2014, from http://www.hl7.org.uk/version3group/cda.asp.

Hopf, Y. M., C. Bond, et al. (2014). "Views of healthcare professionals to linkage of routinely collected healthcare data: a systematic literature review." Journal of the American Medical Informatics Association **21**(e1): e6-e10.

Hughes, B. (2005). Metadata quality evaluation: Experience from the open language archives community. Digital Libraries: International Collaboration and Cross-Fertilization, Springer**:** 320-329.

Hyde, M. K. and K. M. White (2010). "Are organ donation communication decisions reasoned or reactive? A test of the utility of an augmented theory of planned behaviour with the prototype/willingness model." Br J Health Psychol **15**(Pt 2): 435-452.

Ibrahim, T., S. M. Ong, et al. (2004). "The new consent form: is it any better?" Ann R Coll Surg Engl **86**(3): 206-209.

International Health Terminology Standards Development Organisation. "About SNOMED CT." Retrieved January 31, 2014, from http://www.ihtsdo.org/snomed-ct/snomed-ct0/.

International Health Terminology Standards Development Organisation (2013). "SNOMED-CT." Retrieved November 28, 2013, from http://www.ihtsdo.org/snomed-ct/.

International Health Terminology Standards Development Organisation (2014). SNOMED CT Starter Guide, IHTSDO.

Ioannidis, J. A. (2012). "The importance of potential studies that have not existed and registration of observational data sets." JAMA **308**(6): 575-576.

ISO/IEC (2004). Information technology - Metadata registries (MDR) - Part 1: Framework. Switzerland, ISO copyright office.

ISO/IEC (2012). Information technology - Learning, education and training - Metadata for learnng resources Part 2: Dublin Core elements. UK, BSI Standards Limited 2013.

Issa, M. M., E. Setzer, et al. (2006). "Informed Versus Uninformed Consent for Prostate Surgery: The Value of Electronic Consents." Journal of Urology **176**(2): 694-699.

Jenkins, S. P., L. Cappellari, et al. (2006). "Patterns of Consent: Evidence from a General Household Survey." Journal of the Royal Statistical Society. Series A (Statistics in Society) **169**(4): 701-722.

Johnson, S. B., G. Whitney, et al. (2010). "Using global unique identifiers to link autism collections." Journal of the American Medical Informatics Association **17**(6): 689-695.

Johnstone, C. and G. McCartney (2010). "A patient survey assessing the awareness and acceptability of the emergency care summary and its consent model in Scotland." Perspect Health Inf Manag **7**: 1e.

Jonquet, C., N. H. Shah, et al. (2009). "The open biomedical annotator." Summit on Translat Bioinforma **2009**: 56-60.

Kahn, M. G. and C. Weng (2012). "Clinical research informatics: a conceptual perspective." Journal of the American Medical Informatics Association **19**(e1): e36-e42.

Kalra, D. (2006). "Electronic Health Record Standards." IMIA Yearbook 2006: Assessing Information - Technologies for Health **1**(1): 136-144.

Kalton, G. (2012). "Measuring health in population surveys." Statistical Journal of the IAOS: Journal of the International Association for Official Statistics **28**(1): 13-24.

Katayama, T., M. D. Wilkinson, et al. (2013). "The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies." J Biomed Semantics **4**(1): 6.

Keusch, F. (2014). "The Influence of Answer Box Format on Response Behavior on List-Style Open-Ended Questions." Journal of Survey Statistics and Methodology **2**(3): 305-322.

Khan, S. K., K. Karuppaiah, et al. (2012). "The influence of process and patient factors on the recall of consent information in mentally competent patients undergoing surgery for neck of femur fractures." Ann R Coll Surg Engl **94**(5): 308-312.

Klann, J. G., M. D. Buck, et al. (2014). "Query Health: standards-based, cross-platform population health surveillance." Journal of the American Medical Informatics Association.

Klassen, A. F., S. K. Lee, et al. (2005). "Linking survey data with administrative health information: characteristics associated with consent from a neonatal intensive care unit follow-up study." Can J Public Health **96**(2): 151-154.

Knies, G., J. Burton, et al. (2012). "Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population." BMC Health Services Research **12**(1): 52.

Kolker, E., V. Ozdemir, et al. (2014). "Toward more transparent and reproducible omics studies through a common metadata checklist and data publications." Omics **18**(1): 10-14.

Kolker, E. and E. Stewart (2014). "OMICS Studies: How about Metadata Checklist and Data Publications?" J Proteome Res.

Kuchinke, W., C. Ohmann, et al. (2014). "A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model." International Journal of Medical Informatics **83**(12): 941-957.

Kuper, A., S. Reeves, et al. (2008). "An introduction to reading and appraising qualitative research." BMJ **337**: a288.

Kuper, H., A. Nicholson, et al. (2006). "Searching for observational studies: what does citation tracking add to PubMed? A case study in depression and coronary heart disease." BMC Med Res Methodol **6**: 4.

Kush, R. and M. Goldman (2014). "Fostering Responsible Data Sharing through Standards." New England Journal of Medicine **370**(23): 2163-2165.

Kush, R. D., E. Helton, et al. (2008). "Electronic Health Records, Medical Research, and the Tower of Babel." New England Journal of Medicine **358**(16): 1738-1740.

Lavelle-Jones, C., D. J. Byrne, et al. (1993). "Factors affecting quality of informed consent." BMJ **306**(6882): 885-890.

Lee, E. S., R. A. Black, et al. (2015). "Characterizing Secondary Use of Clinical Data." AMIA Jt Summits Transl Sci Proc **2015**: 92-96.

Li, Z., J. Wen, et al. (2012). "ClinData Express--a metadata driven clinical research data management system for secondary use of clinical data." AMIA Annu Symp Proc **2012**: 552-557.

Liolios, K., L. Schrimi, et al. “The Metadata Coverage Index (MCI): A standardized metric for quantifying database metadata richness.” Stand Genomic Sci. 2012 Jul 30;6(3): 438-447

LOINC (2013). "Logical Observation Identifiers Names and Codes ". Retrieved December 02, 2013, from http://loinc.org/.

Luciano, J. S., B. Andersson, et al. (2011). "The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside." J Biomed Semantics **2 Suppl 2**: S1.

Lyons, R. A., D. V. Ford, et al. "Use of data linkage to measure the population health effect of non-health-care interventions." The Lancet **383**(9927): 1517-1519.

Machado, C. M., D. Rebholz-Schuhmann, et al. (2013). "The semantic web in translational medicine: current applications and future directions." Brief Bioinform.

MacKenzie, S. L., M. C. Wyatt, et al. (2012). "Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey." J Am Med Inform Assoc **19**(e1): e119-124.

Marc, D., J. Beattie, et al. “Assessing Metadata Quality of a Federally Sponsored Health Data Repository.” AMIA Annu Symp Proc. 2017; 2016: 864-873

Margaritopoulos, T., M. Margaritopoulos, et al. (2008). A Conceptual Framework for Metadata Quality Assessment.

Mark, J. S. and H. Spiro (1990). "Informed consent for colonoscopy. A prospective study." Arch Intern Med **150**(4): 777-780.

Martin, G. (2003). "Why clinical information standards matter." BMJ **326**.

Matsui, K., R. K. Lie, et al. (2012). "A randomized controlled trial of short and standard-length consent forms for a genetic cohort study: is longer better?" J Epidemiol **22**(4): 308-316.

May, T. P., J. M. P. Craig, et al. (2007). "Viewpoint: IRBs, Hospital Ethics Committees, and the Need for "Translational Informed Consent". [Miscellaneous]." Academic Medicine July **82**(7): 670-674.

Mays, N. and C. Pope (2000). "Qualitative research in health care. Assessing quality in qualitative research." BMJ **320**(7226): 50-52.

McCay, C., J. Evans, et al. (2008). Harmonizing Standards Initiatives: An Overview of Collaborative Standards Initiatives for Clinical Research and Healthcare. European Interchange Tutorial Copenghagen.

McGuire, A. L., J. M. Oliver, et al. (2011). "To share or not to share: a randomized trial of consent for data sharing in genome research." Genet Med **13**(11): 948-955.

McMahon, C., T. Castillo, et al. (2015). "Improving metadata quality assessment in public health and epidemiology." Stud Health Technol Inform **210**: 939.

McMahon, C., S. Denaxas. (2017). “A novel metadata management model to capture consent for record linkage in longitudinal studies”. *in press*.

McMahon, C. and S. Denaxas. "A novel framework for assessing metadata quality in epidemiological and public health research settings.” *AMIA Summits on Translational Science Proceedings*. 2016;2016:199-208.

Meredith, D., S. Crouch, et al. (2010). "Towards a scalable, open-standards service for brokering cross-protocol data transfers across multiple sources and sinks."

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **368**(1926): 4115-4131.

Miller, S. (2011). Metadata for digital collections A how-to-do-it manual London, Facet Publishing.

Min, H., R. Ohira, et al. (2014). "Sharing behavioral data through a grid infrastructure using data standards." Journal of the American Medical Informatics Association **21**(4): 642-649.

Mindell, J., J. P. Biddulph, et al. (2012). "Cohort Profile: The Health Survey for England." Int J Epidemiol **41**(6): 1585-1593.

Moen, W. E., E. L. Stewart, et al. (1998). Assessing metadata quality: findings and methodological considerations from an evaluation of the US Government Information Locator Service (GILS). Research and Technology Advances in Digital Libraries, Santa Barbara.

Moher, D., P. Glasziou, et al. "Increasing value and reducing waste in biomedical research: who's listening?" The Lancet. 2016 Apr 9;387(10027):1573-86. doi: 10.1016/S0140-6736(15)00307-4. Epub

2015 Sep 27.

Moher, D., A. Liberati, et al. (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." PLoS Med **6**(7): e1000097.

Moreira, B. L., M. A. Gonçalves, et al. (2009). "Automatic evaluation of digital libraries with 5SQual." Journal of Informetrics **3**(2): 102-123.

Moreno-Conde, A., D. Moner, et al. (2015). "Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis." J Am Med Inform Assoc **22**(4): 925-934.

Moseley, T. H., M. N. Wiggins, et al. (2006). "Effects of presentation method on the understanding of informed consent." Br J Ophthalmol **90**(8): 990-993.

Mougin, F., A. Burgun, et al. (2006). "Mapping data elements to terminological resources for integrating biomedical data sources." BMC Bioinformatics **7 Suppl 3**: S6.

MRC (2014). Maximising the value of UK population cohorts MRC Strategic Review of the Largest UK Population Cohort Studies. UK.

Musen, M. A., C. A. Bean, et al. (2015). "The center for expanded data annotation and retrieval." Journal of the American Medical Informatics Association **22**(6): 1148-1152.

National Cancer Institute (2014). "NCI Term Browser." Retrieved February 25, 2014, from http://nciterms.nci.nih.gov/ncitbrowser/pages/multiple_search.jsf?nav_type=terminologies.

National Research Ethics Service (2011). Information Sheets & Consent Forms Guidance For Researchers & Reviewers, NHS.

Nature Publishing Group (2014). "Data deposition policies." Retrieved May 09, 2014, from http://www.nature.com/sdata/data-policies/repositories.

Nemeroff, C. B., D. Weinberger, et al. (2013). "DSM-5: a collection of psychiatrist views on the changes, controversies, and future directions." BMC Med **11**: 202.

Neu, S. C., K. L. Crawford, et al. (2012). "Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories." Front Neuroinform **6**: 8.

NIGB (2013). "Section 251." Retrieved February 28, 2013, from http://www.nigb.nhs.uk/s251.

Noy, N. F., N. H. Shah, et al. (2009). "BioPortal: ontologies and integrated data resources at the click of a mouse." Nucleic Acids Res **37**(Web Server issue): W170-173.

Ochoa, X. and E. Duval (2009). "Automatic evaluation of metadata quality in digital repositories." International journal on digital libraries **10**(2-3): 67-91.

Ohno-Machado, L. (2013). "Game changer: how informatics moved from a supporting role to a central position in healthcare." Journal of the American Medical Informatics Association **20**(e2): e197.

Oosthuizen, J. C., P. Burns, et al. (2012). "The changing face of informed surgical consent." J Laryngol Otol **126**(3): 236-239.

Panahiazar, M., M. Dumontier, O. Gevaert. “Predicting Biomedical Metadata in CEDAR: a study of Gene Expression Omnibus.” Journal of Biomedical Informatics. 2017 June 16. ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2017.06.017

Papatheodorou, I., C. Crichton, et al. (2009). "A metadata approach for clinical data management in translational genomics studies in breast cancer." BMC Med Genomics **2**: 66.

Parekh, R., R. Armananzas, et al. (2015). "The importance of metadata to assess information content in digital reconstructions of neuronal morphology." Cell Tissue Res **360**(1): 121-127.

Paris, A., D. Nogueira da Gama Chaves, et al. (2007). "Improvement of the comprehension of written information given to healthy volunteers in biomedical research: a single-blind randomized controlled study." Fundam Clin Pharmacol **21**(2): 207-214.

Pathak, J., K. R. Bailey, et al. (2013). "Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium." Journal of the American Medical Informatics Association **20**(e2): e341-e348.

Pathak, J., R. C. Kiefer, et al. (2012a). "Applying semantic web technologies for phenome-wide scan using an electronic health record linked Biobank." J Biomed Semantics **3**(1): 10.

Pathak, J., R. C. Kiefer, et al. (2012b). "Mining the human phenome using semantic web technologies: a case study for Type 2 Diabetes." AMIA Annu Symp Proc **2012**: 699-708.

Pathak, J., R. C. Kiefer, et al. (2012c). "Using semantic web technologies for cohort identification from electronic health records for clinical research." AMIA Summits Transl Sci Proc **2012**: 10-19.

Pathak, J., J. Wang, et al. (2011). "Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience." Journal of the American Medical Informatics Association **18**(4): 376.

Pereira, S. P., S. H. Hussaini, et al. (1995). "Informed consent for upper gastrointestinal endoscopy." Gut **37**(1): 151-153.

Pesudovs, K., C. K. Luscombe, et al. (2006). "Recall from informed consent counselling for cataract surgery." J Law Med **13**(4): 496-504.

Phimister, E. G. (2015). "Curating the Way to Better Determinants of Genetic Risk." New England Journal of Medicine **372**(23): 2227-2228.

Pickett, B. E., M. Liu, et al. (2013). "Metadata-driven comparative analysis tool for sequences (meta-CATS): An automated process for identifying significant sequence variations that correlate with virus attributes." Virology **447**(1–2): 45-51.

Pilegaard, M. and H. B. Ravn (2012). "Readability of patient information can be improved." Dan Med J **59**(5): A4408.

Pisani, E. and C. AbouZahr (2010). "Sharing health data: good intentions are not enough." Bulletin of the World health Organization **88**(6): 462-466.

Pisani, E., J. Whitworth, et al. (2009). "Time for fair trade in research data." The Lancet **375**(9716): 703-705.

Piwowar, H. A., R. S. Day, et al. (2007). "Sharing detailed research data is associated with increased citation rate." PLoS ONE **2**(3): e308.

Platt, R., R. M. Carnahan, et al. (2012). "The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction." Pharmacoepidemiol Drug Saf **21 Suppl 1**: 1-8.

Pless, B., B. Hagel, et al. (2011). "Different approaches to obtaining consent for follow-up result in biased samples." Inj Prev **17**(3): 195-200.

Pollock, J. and R. Hodgson (2004). Adaptive Information Improving Business Through Semantic Interoperability, Grid Commputing, and Enterprise Integration. Hoboken, New Jersey, John Wiley & Sons, Inc.

Post, L. J., M. Roos, et al. (2007). "A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data." Bioinformatics **23**(22): 3080-3087.

Rahman, L., J. Clamp, et al. (2011). "Is consent for hip fracture surgery for older people adequate? The case for pre-printed consent forms." J Med Ethics **37**(3): 187-189.

Rahmouni, H. B., T. Solomonides, et al. (2010). "Privacy compliance and enforcement on European healthgrids: an approach through ontology." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **368**(1926): 4057-4072.

Raine, R., W. Wong, et al. (2010). "Social variations in access to hospital care for patients with colorectal, breast, and lung cancer between 1999 and 2006: retrospective analysis of hospital episode statistics." BMJ **340**: b5479.

Rans, J., M. Day, et al. (2013). Enabling the citations of datasets generated through public health research. London, Wellcome Trust.

RCUK (2013, 2013 March 6). "RCUK Policy on Open Access." Retrieved April 04, 2013, from http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicyandRevisedguidance.pdf.

RCUK (2015a). "Guidance on best practice in the management of research data." Retrieved June 20, 2016, from http://www.rcuk.ac.uk/documents/documents/rcukcommonprinciplesondatapolicy-pdf/.

RCUK (2015b, 2015 July). "RCUK Common Principles on Data Policy." Retrieved June 20, 2016, from http://www.rcuk.ac.uk/research/datapolicy/.

RCUK (2016). "Concordat on Open Research Data." Retrieved September 15, 2016, from http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/.

Richesson, R. L. and J. Krischer (2007). "Data standards in clinical research: gaps, overlaps, challenges and future directions." J Am Med Inform Assoc **14**(6): 687-696.

Ries, N. M., J. LeGrandeur, et al. (2010). "Handling ethical, legal and social issues in birth cohort studies involving genetic research: responses from studies in six countries." BMC medical ethics **11**(1): 4.

Rogers, C. G., J. E. Tyson, et al. (1998). "Conventional consent with opting in versus simplified consent with opting out: an exploratory trial for studies that do not increase patient risk." J Pediatr **132**(4): 606-611.

Rothwell, E., B. Wong, et al. (2014). "A randomized controlled trial of an electronic informed consent process." J Empir Res Hum Res Ethics **9**(5): 1-7.

Ruan, C. and V. Varadharajan (2003). Supporting e-consent on health data by logic. Foundations of Intelligent Systems. N. Zhong, Z. W. Ras, S. Tsumoto and E. Suzuki. **2871:** 392-396.

Rubbo, B., N. K. Fitzpatrick, et al. (2015). "Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations." International Journal of Cardiology **187**: 705-711.

Safran, C., M. Bloomrosen, et al. (2007). "Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper." J Am Med Inform Assoc **14**(1): 1-9.

Sagotsky, J. A., L. Zhang, et al. (2008). "Life Sciences and the web: a new era for collaboration." Mol Syst Biol **4**: 201.

Samwald, M., A. Coulet, et al. (2012). "Semantically enabling pharmacogenomic data for the realization of personalized medicine." Pharmacogenomics **13**(2): 201-212.

Sbaraini, A., S. M. Carter, et al. (2011). "How to do a grounded theory study: a worked example of a study of dental practices." BMC Med Res Methodol **11**: 128.

Schleyer, T. K. L. and J. L. Forrest (2000). "Methods for the Design and Administration of Web-based Surveys." Journal of the American Medical Informatics Association **7**(4): 416-425.

Schmidt, M. K., E. Vermeulen, et al. (2009). "Regulatory aspects of genetic research with residual human tissue: effective and efficient data coding." Eur J Cancer **45**(13): 2376-2382.

Schweiger, R., S. Hoelzer, et al. (2002). "Plug-and-play XML: a health care perspective." J Am Med Inform Assoc **9**(1): 37-48.

Semantic Web Health Care and Life Sciences Interest Group (2011, 2011 October 18). "Semantic Web Health Care and Life Sciences Interest Group Charter." Retrieved March 13, 2013, from http://www.w3.org/2011/09/HCLSIGCharter#lisci.

Sheehan, M. (2011). "Can broad consent be informed consent?" Public Health Ethics **4**(3): 226-235.

Shivade, C., P. Raghavan, et al. (2013). "A review of approaches to identifying patient phenotype cohorts using electronic health records." J Am Med Inform Assoc.

Shivade, C., P. Raghavan, et al. (2014). "A review of approaches to identifying patient phenotype cohorts using electronic health records." Journal of the American Medical Informatics Association **21**(2): 221-230.

Simborg, D. W., D. E. Detmer, et al. (2013). "The wave has finally broken: now what?" Journal of the American Medical Informatics Association **20**(e1): e21-e25.

Sinaci, A. A. and G. B. Laleci Erturkmen (2013). "A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains." Journal of Biomedical Informatics **46**(5): 784-794.

Singh, R., A. Singh, et al. (2013). "Improvement of workflow and processes to ease and enrich meaningful use of health information technology." Adv Med Educ Pract **4**: 231-236.

Singleton, P. and M. Wadsworth (2006). "Consent for the use of personal medical data in research." British Medical Journal **333**(7561).

Smith, B., M. Ashburner, et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nat Biotech **25**(11): 1251-1255.

Smith, B. and W. Ceusters (2006). "HL7 RIM: an incoherent standard." Stud Health Technol Inform **124**: 133-138.

Song, T. M., H. A. Park, et al. (2014). "Development of health information search engine based on metadata and ontology." Healthc Inform Res **20**(2): 88-98.

Spriggs, M. (2010). "Ethical difficulties with consent in research involving children: Findings from key informant interviews." AJOB Primary Research **1**(1): 34-43.

Stanfill, M. H., M. Williams, et al. (2010). "A systematic literature review of automated clinical coding and classification systems." Journal of the American Medical Informatics Association **17**(6): 646-651.

Steinsbekk, K. S., B. Kåre Myskja, et al. (2013). "Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem?" European Journal of Human Genetics.

Steptoe, A., E. Breeze, et al. (2013). "Cohort Profile: The English Longitudinal Study of Ageing." Int J Epidemiol **42**(6): 1640-1648.

Stvilia, B., L. Gasser, et al. (2007). "A framework for information quality assessment." Journal of the American Society for Information Science and Technology **58**(12): 1720-1733.

Subramanian, S. L., R. R. Kitchen, et al. (2015). "Integration of extracellular RNA profiling data using metadata, biomedical ontologies and Linked Data technologies." J Extracell Vesicles **4**: 27497.

Svanstrom, R., S. Andersson, et al. (2016). "Moving from theory to practice: experience of implementing a learning supporting model designed to increase patient involvement and autonomy in care." BMC Res Notes **9**: 361.

Swensen, S. J., G. S. Meyer, et al. (2010). "Cottage Industry to Postindustrial Care — The Revolution in Health Care Delivery." New England Journal of Medicine **362**(5): e12.

Tao, C., D. Song, et al. (2013). "Semantator: Semantic annotator for converting biomedical text to linked data." Journal of Biomedical Informatics **46**(5): 882-893.

Tate, A. R., L. Calderwood, et al. (2006). "Mother's consent to linkage of survey data with her child's birth records in a multi-ethnic national cohort study." International Journal of Epidemiology **35**(2): 294-298.

Taylor, C. F., D. Field, et al. (2008). "Promoting coherent minimum reporting guidelines for biological and biomedical investigations : the MIBBI project." Nature biotechnology **26**(8): 889-896.

Tenenbaum, J. D., S.-A. Sansone, et al. (2013). "A sea of standards for omics data: sink or swim?" Journal of the American Medical Informatics Association.

Tenopir, C., S. Allard, et al. (2011). "Data sharing by scientists: practices and perceptions." PLoS ONE **6**(6): e21101.

Terry, N. P. and L. P. Francis (2007). "Ensuring the privacy and confidentiality of electronic health records." University of Illinois Law Review(2): 681-735.

The British Standards Institution (2013). BS EN ISO 13120:2013 Health informatics. Syntax to represent the content of healthcare classification systems. Classification Markup Language (ClaML). UK, BSI Standards LImited.

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

Thomassen, J. P., K. Ahaus, et al. (2014). "Developing and implementing a service charter for an integrated regional stroke service: an exploratory case study." BMC Health Serv Res **14**: 141.

Thornton, J. (2015). "What you need to know to make the most of big data in biology." The Lancet **385, Supplement 1**: S5-S6.

Timimi, H., D. Falzon, et al. (2012). "WHO guidance on electronic systems to manage data for tuberculosis care and control." J Am Med Inform Assoc **19**(6): 939-941.

Tirado-Ramos, A., J. Hu, et al. (2002). "Information object definition-based unified modeling language representation of DICOM structured reporting: a case study of transcoding DICOM to XML." J Am Med Inform Assoc **9**(1): 63-71.

Toga, A. W., I. Foster, et al. (2015). "Big biomedical data as the key resource for discovery science." Journal of the American Medical Informatics Association **22**(6): 1126-1131.

U. S. National Library of Medicine (2015, 2015 June 18). "Medical Subject Headings ". Retrieved September 22, 2015, from https://www.nlm.nih.gov/mesh/.

United States National Library of Medicine (2013a, 2013 August 30). "Medical Subject Headings." Retrieved January 30, 2014, from http://www.nlm.nih.gov/mesh/.

United States National Library of Medicine (2013b, 2013 July 24). "MeSH Tree Structure." Retrieved January 31, 2014, from http://www.nlm.nih.gov/mesh/intro_trees.html.

Van den Eynden, V. (2012). "Maximisation of the value of population health sciences data." The Lancet **380, Supplement 3**(0): S76.

Vardaki, M., H. Papageorgiou, et al. (2009). "A statistical metadata model for clinical trials' data management." Computer Methods and Programs in Biomedicine **95**(2): 129-145.

Vawdrey, D. K., C. Weng, et al. (2014). "Enhancing electronic health records to support clinical research." AMIA Jt Summits Transl Sci Proc **2014**: 102-108.

Vreeman, D. (2013, 2013 December 27). "LOINC Version 2.46 abd RELMA Version 6.4 Available." Retrieved February 04, 2014, from http://loinc.org/news/loinc-version-2-46-and-relma-version-6-4-available.html/.

W3C (2012). "OWL 2 Web Ontology Language Document Overview (Second Edition)." Retrieved November 25, 2013, from http:/http://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

W3C (2013). "Semantic Web Health Care and LIfe Sciences (HCLS) Interest Group." Retrieved November 25, 2013, from http://www.w3.org/blog/hcls/.

W3C (2014a, 2014 February 1). "ConverterToRDF." Retrieved February 25, 2014, from http://www.w3.org/wiki/ConverterToRdf.

W3C (2014b). "W3C." Retrieved June 17, 2014, from http://www.w3.org/.

Walport, M. and P. Brest (2011). "Sharing research data to improve public health." The Lancet **377**(9765): 537-539.

Wang, F., C. Vergara-Niedermayr, et al. (2014). "Metadata based management and sharing of distributed biomedical data." Int J Metadata Semant Ontol **9**(1): 42-57.

Wang, X., R. Gorlitsky, et al. (2005). "From XML to RDF: how semantic web technologies will change the design of 'omic' standards." Nat Biotech **23**(9): 1099-1103.

Warner, D. (2011). "HIE patient consent model options." Journal of AHIMA / American Health Information Management Association **82**(5): 48-49.

Warodomwichit, D., D. K. Arnett, et al. (2009). "Polyunsaturated fatty acids modulate the effect of TCF7L2 gene variants on postprandial lipemia." J Nutr **139**(3): 439-446.

Weber, G. M., K. D. Mandl, et al. (2014). "Finding the missing link for big biomedical data." JAMA.

Whetzel, P. L. (2013). "NCBO Technology: Powering semantically aware applications." J Biomed Semantics **4 Suppl 1**: S8.

Whitehead, L. (2011). "Methodological issues in Internet-mediated research: a randomized comparison of internet versus mailed questionnaires." J Med Internet Res **13**(4): e109.

Wilkinson, M. D., M. Dumontier, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." Scientific Data **3**: 160018.

Williams, C. L., K. J. Bunch, et al. (2013). "Cancer Risk among Children Born after Assisted Conception." New England Journal of Medicine **369**(19): 1819-1827.

Wilson, A. J. (2007). "Toward Releasing the Metadata Bottleneck : A Baseline Evaluation of Contributor-Supplied Metadata." Library resources & technical services **51**(1): 16-29.

Winkelman, W. J., K. J. Leonard, et al. (2005). "Patient-perceived usefulness of online electronic medical records: employing grounded theory in the development of information and communication technologies for use by patients living with chronic illness." J Am Med Inform Assoc **12**(3): 306-314.

Witt, C. M., D. Pach, et al. (2009). "Safety of acupuncture: results of a prospective observational study with 229,230 patients and introduction of a medical information and consent form." Forsch Komplementmed **16**(2): 91-97.

World Health Organization (2007). Everybody's business: Strengthening health systems to improve health outcomes: WHO's framework for action. Geneva.

World Health Organization (2012). Electronic recording and reporting for tuberculosis care and control. Geneva.

World Health Organization (2013a). "International Classification of Diseases." Retrieved November 11, 2013, from http://www.who.int/classifications/icd/en/.

World Health Organization (2013b). Research for universal health coverage: World health report 2013. Geneva.

World Health Organization (2014a). "FAQ on ICD." Retrieved January 31, 2014, from http://www.who.int/classifications/help/icdfaq/en/index.html.

World Health Organization (2014b). "The WHO Family of International Classifications." Retrieved January 31, 2014, from http://www.who.int/classifications/en/.

Wyatt, J. C. (2000). "When to Use Web-based Surveys." Journal of the American Medical Informatics Association **7**(4): 426-430.

Zeng, M. and J. Qin (2008). Metadata. New York, Neal-Schuman Publishers, Inc.

Zia, M. I., R. Heslegrave, et al. (2011). "Post-trial period surveillance for randomised controlled cardiovascular studies: submitted protocols, consent forms and the role of the ethics board." J Med Ethics **37**(12): 762-765.

Glossary

ABPI	Association of the British Pharmaceutical Industry
ALSPAC	Avon Longitudinal Study of Parents and Children
BRIDG	Biomedical Research Integrated Domain Group
caBIG	Cancer BioInformatics Grid
CDA	Clinical Document Architecture
CDISC	Clinical Data Interchange Standards Consortium
CHERRIES	Checklist for Reporting Results of Internet E-Surveys
CLSA	Canadian Longitudinal Study of Aging, Étude longitudinale canadienne sur le vieillissement
CPRD	The Clinical Practice Research Datalink
CRI	Clinical Research Informatics
DC	Dublin Core
DCC	Digital Curation Centre
DCMI	Dublin Core Metadata Initiative
DDI	Data Documentation Initiative
DFG	Deutsche Forschungsgemeinschaft
DMP	Data Management Plan
DNBC	Danish National Birth Cohort study
DSM	Diagnostic and Statistical Manual of Mental Disorders
EA	Enterprise Architect
EHR(s)	Electronic Health Record(s)
ELSA	English Longitudinal Study of Ageing
ELSST	European Language Social Science Thesaurus
eMERGE	Electronic Medical Records and Genomics
ESRC	Economic and Social Research Council
FHIR	Fast Health Interoperable Resources
HASSET	Humanities and Social Science Electronic Thesaurus
HCLS IG	World Wide Web Consortium Semantic Web Health Care and Life Sciences Interest Group
HES	Hospital Episode Statistics
HIPAA	The Health Insurance Portability and Accountability Act
HITECH	Health Information Technology for Economic and Clinical Health Act
HL7	Health Level 7
HSCIC	The Health and Social Care Information Centre

ICD	International Classification of Disease
ICPSR	Inter-university Consortium for Political and Social Research
ISO	International Organization for Standardization
LOINC	Logical Observation Identifiers Names and Codes
LSAC	The Longitudinal Study of Australian Children
LS DAM	Life Sciences Domain Analysis Model
MCS	Millennium Cohort Study
MeSH	Medical Subject Headings
MIDUS	Midlife in the United States
NHS	National Health Service
NIH	National Institutes of Health
NLP	Natural Language Processing
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBO Foundry	Open Biomedical Ontologies Foundry
ODM	Operational Data Model
OO	Object Orientation /Object oriented
OWL	Web Ontology Language
PDF	Portable Document Format
RDF	Resource Description Framework
RDL	Research data lifecycle
REDCap	Research Electronic Data Capture
REF	Research Excellence Framework
RIM	Reference Information Model
SDMX	Statistical Data and Metadata Exchange
SDMX HD	Statistical Data and Metadata Exchange Health Domain
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SPARQL	Simple Protocol and RDF Query Language
SWAN	Semantic Web Applications in Neuromedicine
SWT(s)	Semantic Web technology(ies)
TB	Tuberculosis
TMKB	Translational Medicine Knowledge Base
TMO	Translational Medicine Ontology
UAMS	University of Arkansas for Medical Sciences
UML	Unified Modelling Language

UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WHO	World Health Organization
XML	eXtensible Markup Language
XML DTD	Extensible Markup Language Document Type Definition

Appendix

Appendix A: Enhancing the discoverability of public health and epidemiological research data

Supplementary Table 1 The 49 studies and organisations

ID	Name
1	1958 National Child Development Study (UK)
2	1970 British Cohort Study (UK)
3	ALPHA Network
4	Analysis of a sample of type 2 diabetic patients with obesity or overweight and at cardiovascular risk: a cross sectional study in Spain
5	Australasian Association of Cancer Registries
6	Avon Longitudinal Study of Parents and Children (UK)
7	Birth to Twenty
8	Born in Bradford (UK)
9	Cancer Registries
10	CartaGene
11	Concord 2
12	ELFE, Growing up in France (France)
13	European Prospective Investigation into Cancer and Nutrition (EPIC)
14	European Social Survey
15	Generation Scotland (UK)
16	Growing up in New Zealand
17	ICPSR
18	INDEPTH Network
19	International Epidemiological Databases to Evaluate AIDS (IeDEA) in sub-Saharan Africa
20	IPUMS International Project
21	Longitudinal Study of Young People in England (UK)
22	Measure DHS, Demographic Health Surveys

23 MIDUS Midlife in the US

24 Millennium Cohort Study (UK)

25 Monitoring the efficacy and safety of three artemisinin based-combinations therapies in Senegal: results from two years surveillance

26 Norwegian Mother and Child Cohort Study (Norway)

27 Pelotas Birth Cohort Study

28 Population Health Metrics Research Consortium Gold Standard Verbal Autopsy Data 2005–2011

29 Prevalence of schistosome antibodies with hepatosplenic signs and symptoms among patients from Kaoma, Western Province, Zambia

30 RAND Centre for the Study of Ageing

31 SABE - Survey on Health, Well-Being and Aging in Latin America and the Caribbean

32 SABRE Southall and Brent Revisited

33 Scottish Longitudinal Study (Scotland)

34 Study of Environment on Aboriginal Resilience and Child Health

35 The China Kadoorie Biobank

36 The China-Anhui birth cohort study

37 The Epidemiology - France web portal A collaborative project in epidemiology

38 The International Collaboration of Incident HIV and Hepatitis C in Injecting Cohorts Study

39 The Limache birth cohort study

40 The Motorik-Modul Longitudinal Study (Germany)

41 The National Survey of Sexual Attitudes and Lifestyles (UK)

42 The spectrum of paediatric cardiac disease presenting to an outpatient clinic in Malawi

43 TwinsUK and Healthy Aging Twin Study (UK)

44 Udaipur health and Immunization studies

45	UK Biobank (UK)
46	Understanding Society (UK)
47	Whitehall Study (UK)
48	WHO Study on Global AGEing and Adult Health (SAGE)
49	Worldwide Antimalarial Resistance Network

Copy of the survey

Data Discoverability Survey

Thank you for your interest in completing this survey, which has been commissioned by Wellcome Trust to support the work of the Public Health Research Data Forum.

The aim of this survey is to enhance our understanding of the current challenges facing everyone in making research datasets 'discoverable'. A great deal of emphasis is currently being placed on the promotion of data sharing, however this is only part of the challenge. The following questions seek to explore the issues in more detail but are necessarily brief.

We have tried to design this survey so that it is easy to complete within one sitting of no more than 10 minutes although it is possible for you to save and return if you run out of time.

Thank you again for taking the time.

Background information

Please use the following sections to tell us a bit about yourself.

Main employer or employment status

□ University
□ Government agency
□ Non-governmental organization
□ Charity
□ Private company
□ Self-employed
□ Student
□ Unemployed
□ Retired

(Please select the option that applies best to you)

Please indicate the regions of the world where you carry out your work

□ Southern Asia
□ Eastern Asia
□ Europe
□ South-Eastern Asia
□ South America
□ Eastern Africa
□ Northern America
□ Western Africa
□ Western Asia
□ Northern Africa
□ Central America
□ Middle Africa
□ Central Asia
□ Southern Africa
□ Caribbean
□ Oceania

(Please tick all that apply)

How would you describe your role in public health research data

□ Data provider
□ Data user
□ Archivist / Librarian
□ Funding agency

- □ Policy maker
- □ Observer
- □ Other

(Please select all that apply)

Other role: ________________________________

(Please briefly describe the role)

Are you in receipt of funding from any of the following agencies

- □ Agency for Healthcare Research and Quality (USA)
- □ Bill and Melinda Gates Foundation
- □ Canadian Institutes of Health Research
- □ Centres for Disease Control and Prevention
- □ Deutsche Forschungsgemeinschaft (DFG)
- □ Doris Duke Charitable Foundation
- □ Economic and Social Research Council (UK)
- □ Health Research Council of New Zealand
- □ Health Resources and Services Administration (USA)
- □ Hewlett Foundation
- □ INSERM
- □ Medical Research Council (UK)
- □ National Health and Medical Research Council
- □ (Australia)
- □ National Institutes of Health (USA)
- □ Substance Abuse and Mental Health Services
- □ Administration (USA)
- □ Wellcome Trust
- □ The World Bank
- □ NIHR (UK)
- □ Other(s)

(Please tick all that apply)

Other funders (please list one per line): ________________________________

Please indicate the forms of data that you commonly handle

- □ Survey
- □ Healthcare records
- □ Disease registries
- □ Ethnographic
- □ Geospatial
- □ Environmental
- □ Genomic/Proteomic/Metabolomic
- □ Imaging
- □ Physiological measurement
- □ Other

(Please tick all that apply)

What other form(s) of data do you commonly work with? ________________________________

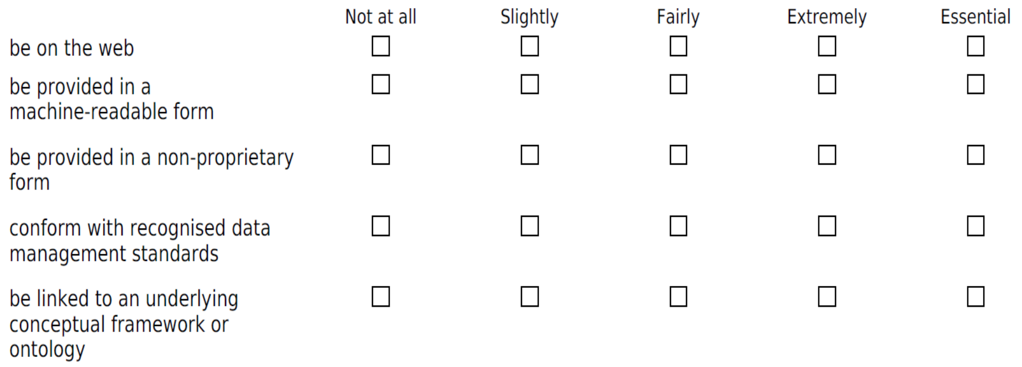
Please indicate which areas of the research data life-cycle are you actively involved in

- □ Conceptualisation
- □ Creation or receipt
- □ Appraisal & Selection
- □ Analysis
- □ Metadata creation
- □ Preservation action
- □ Storage
- □ Access, use and reuse
- □ Transformation
- □ Data Destruction
- □ Archive management
- □ Administration

Data Discoverability

What are the most important things that are needed to promote data discoverability?

Please indicate below how important you consider each aspect of discoverable data:



	Not at all	Slightly	Fairly	Extremely	Essential
be on the web	□	□	□	□	□
be provided in a machine-readable form	□	□	□	□	□
be provided in a non-proprietary form	□	□	□	□	□
conform with recognised data management standards	□	□	□	□	□
be linked to an underlying conceptual framework or ontology	□	□	□	□	□

Preferred search options

- □ Keyword
- □ Subject terms
- □ Concepts
- □ Related concepts

(Please tick all that apply)

What aspects of a research study should ideally be easily searchable?

- □ Research study question
- □ Research study protocol
- □ Research data management plan
- □ Consent form and associated information pack
- □ Funding details
- □ Data collection instrument designs
- □ Variables
- □ Code lists
- □ Concepts
- □ Research publications

(Please tick all that apply)

Is there any other aspect of data discoverability that you feel is important? ________________________________ (Please provide any other observation or concern)

How could discoverability of public health data be improved and do you see any immediate priorities?

____

The following is a list of some repository services and classes of repositories that exist, mostly taken from Nature's website. The web-link gives examples of repositories for many of the categories below. Please indicate the repositories that you have used (data access, deposit or both) in the past or anticipate using in future.

	Already used	Intend to use
Dryad	□	□
Figshare	□	□
ClinicalTrials.gov	□	□
Social Science e.g. ICPSR / UK Data Archive	□	□
DNA protein sequences	□	□
Genetic association & genome variation	□	□
Functional genomics	□	□
Proteomics	□	□
Molecular interactions	□	□
Molecular structure	□	□
Taxonomy & species diversity	□	□
Organism or disease specific resources	□	□
Environmental & geoscience	□	□
Other	□	□

You selected "Other" in the list of repositories above. Please provide details.

____

Controlled Vocabularies and Thesauri

Standardised taxonomies are commonly used approaches to enhance data discovery and facilitate comparison across datasets.

Please indicate which of the following terminologies, classifications, thesauri or metathesauri you are familiar with

□ SNOMED CT
□ OPCS-4
□ International Classification of Disease (ICD)
□ Logical Observation Identifiers Names and Codes (LOINC)
□ Diagnostic and Statistical Manual of Mental Disorders (DSM 5)
□ Read Codes
□ Medical Subject Headings (MeSH)
□ Humanities and Social Sciences Electronic
□ Thesaurus (HASSET)
□ European Language Social Science Thesaurus (ELSST)
□ Unified Medical Language Service (UMLS)
□ Other

(Please tick all that apply)

If your answer to the previous question included "Other" please specify ___________________________________

Which tools do you use to assist with the management of controlled vocabularies? For example, UMLS, Library of Congress, WHO Global Health Observatory indicator registry ___________________________________

Data documentation

In order that researchers can reuse research data meaningfully it is important to ensure that data are provided with detailed descriptors, typically this has been in the form of a code book plus ancillary documentation that describes the processes associated with data collection and any processing that has been carried out.

The following is a list of standards to assist with data documentation. Please indicate which of these you have experience of knowingly using.

□ DC
□ SDMX
□ EAD
□ METS
□ DCAT
□ CKAN
□ eGMS
□ INSPIRE
□ ADMS
□ DDI 2/3

(Please tick all that apply)

Which tools do you use to assist you with data documentation? ___________________________________

Which are the key challenges in creating/using documenting data? ___________________________________

Data Citation and Data Publications

Citation of data is becoming an important tool to promote and track data reuse. Data publication offers a mechanism to promote data citation. The following section explores your views and understanding of these options.

It is possible to publish articles that describe research datasets independently of conventional research publications. Are you familiar with this form of data publication?

□ Yes
□ No

Where did you first hear about data publications?

□ Colleague
□ Journal
□ Conference/workshop
□ Search engine
□ Other

(Please tick all that apply)

If your answer to the previous question included 'other' please specify. ___________________________________

Please indicate which benefit(s) of data citation are most important to you

□ Easier for readers to locate data
□ Proper credit given to data contributors
□ Links between datasets and associated methodology publication provide context for reader
□ Links between datasets and publications describing their use can demonstrate impact.
□ Infrastructure can support long-term reference and reuse
□ Less danger of data plagiarism
□ Promotes professional recognition and rewards
□ Other

(Please tick all that apply)

What other benefits do you see in data citation? ___________________________________

How granular should data citations be:

- □ Dataset collections
- □ Single datasets (or sweep)
- □ Files within datasets
- □ Individual items of data
- □ Other

(Please tick all that apply)

Since you selected other above, please give your suggestion here: ___________________________________

Ideally, how should longitudinal and regularly changing datasets be handled?

- □ New identifier assigned at each update
- □ Publish revisions at regular intervals
- □ Time series data should be published as complete 'snapshots'
- □ Time series data should be published in instalments
- □ All published versions of the datasets must be stored
- □ Other

(Please tick all that apply)

Other mechanism suggested for handling longitudinal and changing datasets: ___________________________________

Which would you say, if any, are the key challenges affecting the widespread adoption of data publications? ___________________________________

Supplementary Table 2 Use of metadata standards by role and stage of the research data lifecycle

Metadata standards		Role in public health						Total
		Data provider	Data user	Archivist / Librarian	Funding agency	Policy maker	Other	
DC	Creation or receipt	0	0	3		1	0	3
	Appraisal & Selection	1	1	4		0	0	4
	Analysis	1	1	3		1	0	3
	Metadata creation	1	1	4		1	0	4
	Preservation action	2	1	4		1	0	4
	Storage	1	1	3		1	0	3
	Access, use and reuse	2	1	6		1	1	7
	Transformation	0	0	1		0	0	1
	Data Destruction	0	0	1		1	0	1
	Archive management	2	1	4		1	0	4
	Administration	1	1	1		0	0	1
	Total	2	1	6		1	1	7
SDMX	Conceptualisation	0	1	0		0	1	1
	Appraisal & Selection	1	2	2		0	1	3
	Analysis	1	2	1		1	1	3
	Metadata creation	1	1	2		0	0	2
	Preservation action	1	1	2		0	0	2

	Storage	1	1	2		0	0	2
	Access, use and reuse	1	2	2		0	1	3
	Transformation	0	1	0		0	1	1
	Archive management	1	2	2		0	1	3
	Administration	1	1	1		0	0	1
	Total	1	2	2		1	1	4
EAD	Appraisal & Selection	0		1				1
	Metadata creation	0		1				1
	Preservation action	1		2				2
	Storage	0		1				1
	Access, use and reuse	1		2				2
	Archive management	1		2				2
	Total	1		2				2
METS	Creation or receipt		0	2		1	0	2
	Appraisal & Selection		0	2		0	0	2
	Analysis		2	1		1	0	3
	Metadata creation		0	4		1	0	4
	Preservation action		0	4		1	0	4
	Storage		0	4		1	0	4
	Access, use and reuse		0	3		1	1	4
	Data Destruction		0	1		1	0	1
	Archive management		0	4		1	0	4
	Total		2	4		1	1	7
DCAT	Creation or receipt			1		1	0	1

	Appraisal & Selection			0		0	1	1
	Analysis			1		1	1	2
	Metadata creation			1		1	1	2
	Preservation action			1		1	0	1
	Storage			1		1	0	1
	Access, use and reuse			1		1	1	2
	Data Destruction			1		1	0	1
	Archive management			1		1	0	1
	Total			1		1	1	2
CKAN	Conceptualisation	1	1	0	1	1		1
	Creation or receipt	1	1	0	1	1		1
	Appraisal & Selection	2	1	1	1	1		2
	Analysis	1	1	0	1	1		1
	Metadata creation	1	1	0	1	1		1
	Preservation action	1	0	1	0	0		1
	Storage	2	1	1	1	1		2
	Access, use and reuse	1	1	0	1	1		1
	Archive management	1	0	1	0	0		1
	Administration	1	1	0	1	1		1
	Total	2	1	1	1	1		2
eGMS	Conceptualisation	1	1					1
	Creation or receipt	1	1					1
	Appraisal & Selection	1	1					1
	Analysis	1	1					1

	Preservation action	1	1					1
	Access, use and reuse	1	1					1
	Archive management	1	1					1
	Total	1	1					1
INSPIRE	Conceptualisation	2	2				0	2
	Creation or receipt	2	2				0	2
	Appraisal & Selection	1	1				1	2
	Analysis	2	4				1	5
	Metadata creation	2	2				1	3
	Preservation action	1	1				0	1
	Storage	1	1				0	1
	Access, use and reuse	2	2				1	3
	Transformation	2	2				0	2
	Data Destruction	1	1				0	1
	Archive management	1	1				0	1
	Administration	1	1				0	1
	Total	2	4				1	5
DDI 2/3	Conceptualisation	11	9	1	2	2	2	12
	Creation or receipt	12	9	3	2	2	2	15
	Appraisal & Selection	13	9	9	2	2	0	19
	Analysis	14	11	5	1	1	2	18
	Metadata creation	17	10	9	2	3	2	23
	Preservation action	11	6	8	1	1	2	15

Storage	15	9	9	2	2	2	20
Access, use and reuse	19	12	12	2	3	3	29
Transformation	10	5	3	0	0	1	11
Data Destruction	1	0	1	0	0	1	2
Archive management	9	3	7	0	0	2	12
Administration	6	5	1	1	1	1	6
Total	19	12	12	2	3	3	29

Percentages and totals are based on respondents.

Appendix B: Improving metadata quality assessment in public health and epidemiological research

Supplementary Table 3 Search strategy

Source	Search terms	Results
ACM Digital Library	"Epidemiology" AND "Metadata" AND "Metadata quality assessment" AND "Metadata quality dimensions" AND "Metadata quality evaluation" AND "Public health" AND "Public health and epidemiology" AND "Quality assessment" AND "Quality evaluation"	0
BioMed Central	"Epidemiology" AND "Metadata" AND "Metadata quality assessment" AND "Metadata quality dimensions" AND "Metadata quality evaluation" AND "Public health" AND "Public health and epidemiology" AND "Quality assessment" AND "Quality evaluation"	0
CINAHL Plus	Epidemiology AND Metadata AND Metadata quality assessment AND Metadata quality dimensions AND Metadata quality evaluation AND Public health AND (Public health and epidemiology) AND Quality assessment AND Quality evaluation	0
The Cochrane Library	epidemiology and metadata and metadata quality assessment and metadata quality dimensions and metadata quality evaluation and public health and public health and epidemiology and quality assessment and quality evaluation (Word variations have been searched)	0
EMBASE	(Epidemiology and Metadata and Metadata quality assessment and Metadata quality dimensions and Metadata quality evaluation and Public health and (Public health and epidemiology) and Quality assessment and Quality evaluation).af.	0
Lecture Notes in Computer Science	"Epidemiology" AND "Metadata" AND "Metadata quality assessment" AND "Metadata quality dimensions" AND "Metadata quality evaluation" AND "Public health" AND "Public health and epidemiology" AND "Quality assessment" AND "Quality evaluation"	253
JSTOR	(((((((Epidemiology) AND (Metadata)) AND (Metadata quality assessment)) AND (Metadata quality dimensions)) AND (Metadata quality evaluation)) AND (Public health)) AND (Public health and epidemiology)) ((Quality assessment) AND (Quality evaluation))	134
PubMed	(((((((("epidemiology"[Subheading] OR "epidemiology"[All Fields] OR "epidemiology"[MeSH Terms]) AND Metadata[All Fields]) AND (Metadata[All Fields] AND quality[All Fields] AND ("Assessment"[Journal] OR "assessment"[All Fields]))) AND (Metadata[All Fields] AND quality[All Fields] AND ("Dimensions (N Y N Y)"[Journal] OR "dimensions"[All Fields] OR "DHS Dimens"[Journal] OR "dimensions"[All Fields]))) AND (Metadata[All Fields] AND quality[All Fields] AND ("evaluation studies"[Publication Type] OR	0

	"evaluation studies as topic"[MeSH Terms] OR "evaluation"[All Fields]))) AND ("public health"[MeSH Terms] OR ("public"[All Fields] AND "health"[All Fields]) OR "public health"[All Fields])) AND (("public health"[MeSH Terms] OR ("public"[All Fields] AND "health"[All Fields]) OR "public health"[All Fields]) AND ("epidemiology"[Subheading] OR "epidemiology"[All Fields] OR "epidemiology"[MeSH Terms]))) AND (Quality[All Fields] AND ("Assessment"[Journal] OR "assessment"[All Fields]))) AND (Quality[All Fields] AND ("evaluation studies"[Publication Type] OR "evaluation studies as topic"[MeSH Terms] OR "evaluation"[All Fields]))	
Scopus	(ALL (epidemiology) AND ALL (metadata) AND ALL (metadata quality assessment) AND ALL (metadata quality dimensions) AND ALL (metadata quality evaluation) AND ALL (public health) AND ALL (public health AND epidemiology) AND ALL (quality assessment) AND ALL (quality evaluation))	16
Web of Science	TOPIC: (Epidemiology) *AND* TOPIC: (Metadata) *AND* TOPIC: (Metadata quality assessment) *AND* TOPIC: (Metadata quality dimensions) *AND* TOPIC: (Metadata quality evaluation) *AND* TOPIC: (Public health) *AND* TOPIC: (Public health and epidemiology) *AND* TOPIC: (Quality assessment) *AND* TOPIC: (Quality evaluation)	25

Copy of the metadata quality survey

Metadata Quality Survey

Thank you for taking the time to complete this survey.

With current drives to increase the creation and availability of metadata, and improve application of metadata standards, this survey aims to examine current issues associated with using metadata and in particular, identify ways through which quality may be enhanced.

This survey forms a part of my work focusing on metadata quality within the public health and epidemiology research domains. My work is funded through a MRC CASE Studentship with AIMES Grid Services CIC.

This survey should not take no longer than 10 minutes to complete, but the option to save your answers and return to the survey later is available.

Once again, thank you.

Christiana McMahon

Demographics

Please select your current location

□ Caribbean
□ Europe
□ Eastern Asia
□ South-Eastern Asia
□ Southern Asia
□ Western Asia
□ Central Asia
□ Northern America
□ South America
□ Central America
□ Northern Africa
□ Eastern Africa
□ Southern Africa
□ Western Africa
□ Middle Africa
□ Oceania

7.

Please select your main employer or employment status

□ Charity
□ Government
□ Non governmental agency
□ Private company
□ Retired
□ Self-employed
□ Student
□ Unemployed
□ University

8.

Please select your role(s) in public health and epidemiology research

□ Archivist / librarian
□ Clinician / clinical advisor
□ Data provider
□ Data user
□ Funding agency
□ Policy maker
□ Observer

- □ Other

(Please select all that apply)

If other, please specify ____________________________________

Metadata

How often do you use metadata?

□ Never
□ Sometimes
□ Regularly
□ Frequently
□ Very frequently
□

Please indicate which types of metadata you have used

□ Administrative
□ Descriptive
□ Microdata
□ Semantic
□ Other

(Please select all that apply)

If other, please specify ___________________________________

In which formats does the metadata you use routinely appear?

□ PDF(s)
□ Spreadsheet(s)
□ Word© document(s)
□ XML
□ RDF
□ HTML
□ Other

(Please select all that apply)

If other, please specify ___________________________________

At which points in the research data lifecycle have you handled metadata?

□ Conceptualisation
□ Creation or receipt

□ Appraisal & Selection
□ Analysis
□ Preservation action
□ Access, use and reuse
□ Transformation
□ Data destruction
□ Archive management
□ Administration

(Please select all that apply)

At which levels should metadata be available?

□ Research study level
□ Single dataset / sweep of data
□ Variable level
□ Each time a change is made to the data
□ Other

(Please select all that apply)

If other, please specify ___________________________________

Which are the main barriers to creating and/or using metadata in biomedical research? ___________________________________

Tools and technologies

How do you select a tool and/or technology?

Funder requirement

Suggestion from colleagues

Standard practice

Other

(Please select all that apply)

If other, please specify ___________________________________

Are the data you receive documented using a clinical terminology?

□ Yes
□ No

9.

If yes, please indicate which of the following clinical terminologies you have come across

□ DSM
□ ICD
□ LOINC
□ MeSH
□ OPCS
□ Read Codes
□ SNOMED CT
□ Other

(Please select all that apply)

If other, please specify ___________________________________

Did you experience any problems when using the clinical terminology(ies)?

□ Yes
□ No

If yes, please briefly describe the problems you encountered ___________________________________

Does the metadata you routinely use comply with any standards?

□ Yes
□ No

10.
11.
12.

If yes, please indicate which of the following metadata standards you have used

□ Dublin Core (or derived standard)
□ Data Documentation Initiative 2 (Code book)
□ Data Documentation Initiative 3 (Lifecycle)
□ ISO/IEC 11179
□ MIBBI (Minimum Information for Biological and Biomedical Investigations)
□ Observ-OM
□ OME-XML (Open Microscopy Environment XML)
□ Protocol Data Element Definitions
□ SDMX / SDMX-HD (Health Domain)
□ Other

(Please select all that apply)

If other, please specify ___________________________________

Have you ever used a metadata catalogue to help improve the discoverability of your research?

□ Yes
□ No

Did you experience any challenges in doing so?

□ Yes
□ No

If yes, please briefly describe the challenges you experienced ______________________________________

Have you ever used a metadata catalogue to identify and characterise research datasets?

□ Yes
□ No

13.

Did you experience any difficulties in accessing and/or using the metadata?

□ Yes
□ No

If yes, please briefly describe the difficulties you experienced ______________________________________

Have you ever used Semantic Web technologies when handling metadata e.g. RDF, OWL etc.

□ Yes
□ No

If yes, did you experience any challenges?

□ Yes
□ No

If yes, please briefly describe the challenges you experienced ____________________________________

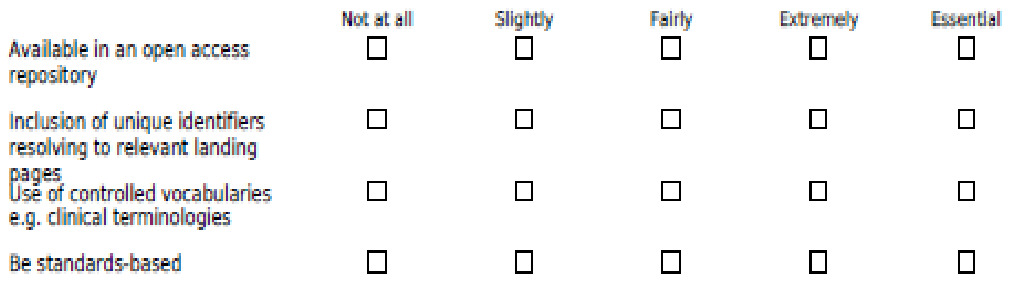
Have you used any kind of metamodel as part of your research e.g. HL7 RIM?

□ Yes
□ No

If yes, please specify which one(s) ____________________________________

Metadata usability

Please indicate which of the following you consider to be of importance to metadata usability



	Not at all	Slightly	Fairly	Extremely	Essential
Available in an open access repository	□	□	□	□	□
Inclusion of unique identifiers resolving to relevant landing pages	□	□	□	□	□
Use of controlled vocabularies e.g. clinical terminologies	□	□	□	□	□
Be standards-based	□	□	□	□	□

Are there any other aspects you would consider to be of importance to metadata usability? ________________________________

Quality Assessment

Which of the following quality dimensions do you feel are of importance to good quality metadata in biomedical research?

- □ Accessibility (extent to which the metadata can be accessed)
- □ Accuracy (correctness of the metadata)
- □ Appropriateness (extent to which the metadata are relevant)
- □ Comprehensiveness (extent to which the metadata are complete)
- □ Discoverability (how visible the metadata are - can it be easily found)
- □ Extendibility (extent to which the metadata may be easily extended)
- □ Interoperability (extent to which metadata can be exchanged and used without problems)
- □ Meta-metadata (metadata about the metadata)
- □ Timeliness (is the metadata current, inclusion of temporal information)
- □ Versionability (extent to which a new version may be easily created)
- □ Other

(Please select all that apply)

If other, please specify ________________________________

How often do you assess the quality of the metadata you routinely handle?

- □ Never
- □ Sometimes
- □ Regularly
- □ Frequently
- □ Very frequently

14.

Do you use any kind of metadata assessment criteria?

- □ Yes
- □ No

If yes, please select one option from the following

- □ Add name of assessment criteria
- □ Add link to web page detailing assessment criteria
- □ Upload file containing assessment criteria

Name of metadata assessment criteria ____________________________________

Link to web page detailing the metadata assessment criteria ____________________________________

Upload file containing method of metadata assessment criteria

What do you see as being the main difficulties when trying to assess metadata quality in biomedical research? ____________________________________

Which are the immediate priorities when addressing metadata quality? ____________________________________

Are you interested in participating in any small group discussions or short interviews regarding metadata quality?

□ Yes
□ No

As this survey is anonymous, please enter your email address if you give consent to be contacted ____________________________________

Supplementary Table 4 Millennium Cohort Study

Area of metadata quality assessment	**Headings**	**Results**
General information	Types of metadata	Descriptive and administrative
	Formats of metadata	PDFs from the Centre of Longitudinal Studies website http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=883&sitesectiontitle=The+age+11+survey+of+the+MCS+%282012%29 XML from UK Data Service metadata catalogue http://esds.ac.uk/DDI25/7464.xml
	Granularity of metadata	Study level (http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=851&sitesectiontitle=Welcome+to+the+Millennium+Cohort+Study) Single dataset/sweep of data (http://discover.ukdataservice.ac.uk/catalogue/?sn=7464&type=Data%20catalogue) Variable level (http://nesstar.ukdataservice.ac.uk/webview/index.jsp?v=2&mode=documentation&submode=abstract&study=http://nesstar.ukdataservice.ac.uk:80/obj/fStudy/7464&top=yes)
	Missing or incomplete metadata	The artefacts reviewed were: Millennium Cohort Study Fifth Sweep (MCS5) Technical Report, UK Data Service data catalogue record for Millennium Cohort Study Fifth Survey 2012 and entry in the Nesstar catalogue were all complete.

Tools and technologies	Structure of metadata E.g. continuous prose, sectioned,	The technical report was a PDF comprising of continuous prose broken down by numbered paragraphs. The UK Data Service record was a list consisting of headings and sub-sections Nesstar catalogue had a series of collapsible headings with each sub-section providing increasing granularity of information
	Presence of clinical terminologies	Could not find any in the randomly selected subset of documentation reviewed.
	Indexing in catalogues	Metadata are indexed in Nesstar and the UK Data Service
	Restrictions on access to metadata	Did not experience any restrictions
	Application of Semantic Web technologies	DDI 2.5 compliant XML format publically available http://esds.ac.uk/DDI25/7464.xml
	Method of application and reason(s) for use	Could not find this information
Usability	Metadata repositories	Nesstar publisher (http://nesstar.ukdataservice.ac.uk/webview/index.jsp?v=2&mode=documentation&submode=abstract&study=http://nesstar.ukdataservice.ac.uk:80/obj/fStudy/7464&top=yes) UK Data Service (http://discover.ukdataservice.ac.uk/catalogue/?sn=7464&type=Data%20catalogue)
	Metadata standards	Data Documentation Initiative

	Cross-walks or other mappings	Could not find cross-walks or other mappings
Management and curation	Creation of metadata	Nesstar catalogue provides authoring entity and identification number the XML syntax provides two agencies and their identification numbers. The XML also provides the production date and to whom the copyright is assigned.
	Versioning	Version information was found in the XML – the date and version number.

Supplementary Table 5 National Survey of Midlife Development in the United States (MIDUS II): Biomarker Project, 2004-2009

Area of metadata quality	**Headings**	**Findings**
General information	Names and locations of research artefacts reviewed	National Survey of Midlife Development in the United States (MIDUS II): Biomarker Project, 2004-2009 metadata record http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/203/studies/29282?archive=ICPSR&amp;sortBy=7 National Survey of Midlife Development in the United States (MIDUS II): Biomarker Project, 2004-2009 DDI Codebook www.icpsr.umich.edu/cgi-bin/file?comp=none&study=29282&ds=1&file_id=1101115&path=ICPSR National Survey of Midlife Development in the United States (MIDUS II): Biomarker Project, 2004-2009 data file notes www.icpsr.umich.edu/cgi-bin/file?comp=none&study=29282&ds=1&file_id=1101124&path=ICPSR
	Types of metadata	Descriptive, administrative, semantic
	Formats of metadata	PDF, XML
	Granularity of metadata	Study http://www.icpsr.umich.edu/icpsrweb/content/membership/or/metadata/index.html Sweep http://www.icpsr.umich.edu/icpsrweb/ICPSR/series/203/studies/29282?archive=ICPSR&amp;sortBy=7
	Missing or incomplete metadata	Complete
	Online data visualisation	Home page states there is ‘online analysis version with question text’ but I could not find this.
	Provision of variable descriptions	Yes - http://www.icpsr.umich.edu/icpsrweb/ICPSR/ssvd/studies/29282/variables

	Links to other studies, sweeps or publications	Links to publications for this sweep http://www.icpsr.umich.edu/icpsrweb/ICPSR/biblio/studies/29282/resources?collection=DATA&archive=ICPSR&sortBy=1 and study http://www.icpsr.umich.edu/icpsrweb/ICPSR/biblio/series/00203/resources?sortBy=1&archive=ICPSR
	Other	
Tools and technologies	Structure of metadata E.g. continuous prose, sectioned,	Continuous prose, signposted with clear headings DDI compliant XML can be found if the page source is viewed http://www.icpsr.umich.edu/icpsrweb/ICPSR/ddi3/studies/29282
	Presence of clinical terminologies	Could not find these
	Presence of code(s) and category(ies) lists	Description of variables coding conventions
	Indexing in catalogues	Indexed in ICPSR
	Restrictions on access to metadata	Did not find any
	Application of Semantic Web technologies	XML
	Method of application and reason(s) for use	Could not find this information
Usability	Metadata standards	DDI Codebook and DDI Lifecycle Dublin Core MARC21 XML Datacite XML

	Cross-walks and other inclusive of method and when these were created	Could not find cross-walks between the different versions of DDI or across the different standards
	Other mappings	
	Provision of metadata model (metamodels)	Could not find this
Management and curation	Date and version of assessment	20150519; version 1
	Name of person assessing the metadata	C McMahon
	Creation of metadata	All dates provided alongside the version history
	Provision of other versions	Full version history is available inclusive of brief description of changes e.g. on 2013-04-13, ‘technical corrections were made to the data formats’.

Supplementary Table 6 Danish National Birth Cohort

Area of metadata quality	**Headings**	**Findings**
General information	Names and locations of research artefacts reviewed	Danish National Birth Cohort http://www.ssi.dk/English/RandD/Research%20areas/Epidemiology/DNBC.aspx Danish National Birth Cohort (Centre of Longitudinal Studies) http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=342&sitesectiontitle=Danish+National+Birth+Cohort The Danish National Birth Cohort – its background, structure and aim http://www.ssi.dk/~/media/Indhold/DK%20-%20dansk/Forskning/BSMB/BSMB%20artikler/danishbirthcohort.ashx
	Types of metadata	Descriptive
	Formats of metadata	PDF, HTML
	Granularity of metadata	Study http://www.ssi.dk/~/media/Indhold/DK%20-%20dansk/Forskning/BSMB/BSMB%20artikler/danishbirthcohort.ashx http://www.ssi.dk/English/RandD/Research%20areas/Epidemiology/DNBC/About%20the%20DNBC/Background%20and%20Overall%20aim%20of%20the%20DNBC.aspx
	Missing or incomplete metadata	Complete
	Online data visualisation/ variable tabulation tools	Found graph showing cumulated participation across the different interviews http://www.ssi.dk/English/RandD/Research%20areas/Epidemiology/DNBC/For%20researchers/Data%20available.aspx Variable description
	Links to other studies,	Through the Danish Data Archive:

	sweeps or publications	Danish National Birth Cohort I, 1997-2003 http://samfund.dda.dk/ddakatalog/sdfiler/R17541gb.htm Danish National Birth Cohort II, 1997-2003 http://samfund.dda.dk/ddakatalog/sdfiler/R22345gb.htm Danish National Birth Cohort III, 1998-2003 http://samfund.dda.dk/ddakatalog/sdfiler/R23163gb.htm Danish National Birth Cohort IV, 1999-2004 http://samfund.dda.dk/ddakatalog/sdfiler/R23164gb.htm List of NCI indexed publications http://www.ncbi.nlm.nih.gov/myncbi/browse/collection/40488075/?sort=date&direction=descending List of DNBC theses http://www.ssi.dk/English/RandD/Research%20areas/Epidemiology/DNBC/Publications/DNBC%20Theses.aspx
	Other	
Tools and technologies	Structure of metadata E.g. continuous prose, sectioned,	Continuous prose sign posted using headings
	Presence of clinical terminologies	Could not find any
	Presence of code(s) and category(ies) lists	Codebooks are available for all four interviews I http://www.ssi.dk/~/media/Indhold/DK%20-%20dansk/Forskning/BSMB/BSMB%20Dokumenter/Kodeboeger/De%20fire%20forste%20interviews/UKInt1.ashx II http://www.ssi.dk/~/media/Indhold/DK%20-%20dansk/Forskning/BSMB/BSMB%20Dokumenter/Kodeboeger/De%20fire%20forste%20interviews/UKInt2.ashx III http://www.ssi.dk/~/media/Indhold/DK%20-%20dansk/Forskning/BSMB/BSMB%20Dokumenter/Kodeboeger/De%20fir

		e%20forste%20interviews/Int3UKJune08.ashx IV http://www.ssi.dk/~/media/Indhold/DK%20-%20dansk/Forskning/BSMB/BSMB%20Dokumenter/Kodeboeger/De%20fire%20forste%20interviews/UKInt4.ashx
	Indexing in catalogues/repositories	Centre for Longitudinal Studies http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=342&sitesectiontitle=Danish+National+Birth+Cohort Birthcohorts.net http://www.birthcohorts.net/bch2/?action=show&UserID=3 The page did not load properly and the information was difficult to read
	Restrictions on access to metadata	No restrictions experienced
	Application of Semantic Web technologies	Description can be found here http://dda.dk/search-technical-information?lang=en DDI can be found by viewing the page source e.g. data collection http://dda.dk/search-technical-information/cv/collectionsituation.dda.dk-1.0.0.cv
	Method of application and reason(s) for use	Could not find this information
Usability	Metadata standards	Metadata at the Danish Data Archive comply with DDI
	Cross-walks inclusive of method and when these were created	Could not find any
	Other mappings	
	Provision of metadata model (metamodels)	Could not find any

Management and curation	Date and version of assessment	20150520; version1
	Name of person assessing the metadata	C McMahon
	Creation of metadata	Dates of when websites were last update were found at the bottom of the screen.
	Provision of other versions	Could not find any

Appendix C: Improving the capture of consent for record linkage metadata in longitudinal studies

Supplementary Table 7 Theme: Analysis

Author	Title
(Hyde and White 2010)	Are organ donation communication decisions reasoned or reactive? A test of the utility of an augmented Theory of Planned Behavior with the Prototype/Willingness Model
(Klassen, Lee et al. 2005)	Linking Survey data with administrative health information characteristics associated with consent from a neonatal intensive care unit follow-up study
(Dunn, Jordan et al. 2004)	Patterns of consent in epidemiologic research: evidence from over 25, 000 responders
(Knies, Burton et al. 2012)	Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population
(Spriggs 2010)	Ethical difficulties With Consent in research Involving Children: Findings from Key Informant Interviews
(Lavelle-Jones, Byrne et al. 1993)	Factors affecting quality of informed consent
(Tate, Calderwood et al. 2006)	Mother's consent to linkage of survey data with her child's birth records in a multi-ethnic national cohort study
(Jenkins, Cappellari et al. 2006)	Patterns of consent: evidence from a General Household Survey
(Khan, Karuppaiah et al. 2012)	The influence of process and patient factors on the recall of consent information in mentally competent patients undergoing surgery for neck of femur fractures
(McGuire, Oliver et al. 2011)	To share or not to share: a randomized trial of consent for data sharing in genome research
(Burns, Magyarody et al. 2011)	Attitudes of the general public toward alternative consent models
(Budin-Ljosne, Tasse et al. 2011)	Bridging consent: from toll bridges to lift bridges?

Supplementary Table 8 Theme: Comparison of models

Author	Title
(Berry, Ryan et al. 2012)	A randomised controlled trial to compare opt-in and opt-out parental consent for childhood vaccine safety surveillance using data linkage
(Mark and Spiro 1990)	Informed consent for colonoscopy a prospective study
(Matsui, Lie et al. 2012)	A Randomized Controlled Trial of Short and Standard-Length Consent Forms for a Genetic Cohort Study: Is Longer Better?
(Berry, Ryan et al. 2011)	A randomised controlled trial to compare opt-in and opt-out parental consent for childhood vaccine safety surveillance using data linkage: study protocol
(Rogers, Tyson et al. 1998)	Conventional consent with opting in versus simplified consent with opting out: An exploratory trial for studies that do not increase patient risk
(Pless, Hagel et al. 2011)	Different approaches to obtaining consent for follow-up result in biased samples
(Pereira, Hussaini et al. 1995)	Informed consent for upper gastrointestinal endoscopy
(Zia, Heslegrave et al. 2011)	Post-trial period surveillance for randomized controlled cardiovascular studies: submitted protocol, consent forms and the role of the ethics board
(Fernando, Bhojwani et al. 2007)	Standards in consent for cataract surgery
(Ibrahim, Ong et al. 2004)	The new consent form: is it any better?
(Warner 2011)	HIE Patient Consent Model Options
(Steinsbekk, Kåre Myskja et al. 2013)	Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem?
(Gefenas, Dranseika et al. 2012)	Turning residual human biological materials into research collections: playing with consent
(May, Craig et al. 2007)	Viewpoint: IRBs, Hospital Ethics committees, and the Need for “Translational Informed Consent”
(Issa, Setzer et al. 2006)	Informed versus uninformed consent for prostate surgery: The value of electronic consents

Supplementary Table 9 Theme: Consent aspects of secondary uses of data

Author	Title
(Elger, Iavindrasana et al. 2010)	Strategies for health data exchange for secondary cross-institutional clinical research
(Singleton and Wadsworth 2006)	Consent for the use of personal medical data in research

Supplementary Table 10 Theme: Development of a new model/form

Author	Title
(Gori, Greco et al. 2012)	A new informed consent form model for cancer patients: Preliminary results of a prospective study by the Italian Association of Medical Oncology (AIOM)
(Davis, Pohlman et al. 2003)	Improving the Process of Informed consent in the Critically Ill
(Gold, Lebel et al. 1993)	Model Consent forms for DNA Linkage Analysis and Storage
(Witt, Pach et al. 2009)	Safety of acupuncture: Results of a prospective observational study with 229,230 patients and introduction of a medical information and consent form
(Caulfield, Upshur et al. 2003)	DNA databanks and consent: A suggested policy option involving an authorization model

Supplementary Table 11 Theme: Development of tools to assist consent process

Author	Title
(Schmidt, Vermeulen et al. 2009)	Regulatory aspects of genetic research with residual human tissue: Effective and efficient data coding
(Coiera and Clarke 2004)	e-Consent: The design and implementation of consumer consent mechanisms in an electronic environment

Supplementary Table 12 Theme: Discussion of a single model of consent

Author	Title
(Johnstone and McCartney 2010)	A patient survey assessing the awareness and acceptability of the emergency care summary and its consent model in Scotland
(Rahman, Clamp et al. 2011)	Is consent for the hip fracture surgery older people adequate? The case for pre-printed consent forms
(Sheehan 2011)	Can broad consent be informed consent?
(Henderson 2011)	Is Informed Consent Broken?

Supplementary Table 13 Theme: Establishing and/or Improving participant understanding

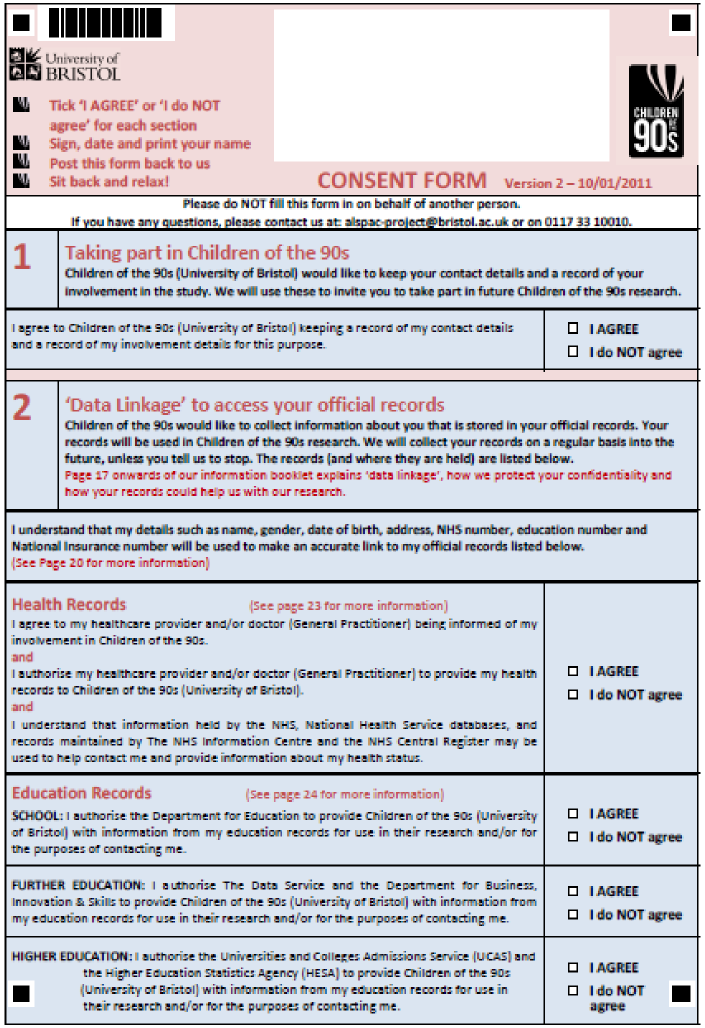
Author	Title
(Friedlander, Loeben et al. 2011)	A novel method to enhance informed consent: a prospective and randomised trial of form-based versus electronic assisted informed consent in paediatric endoscopy
(Arora, Rajagopalan et al. 2011)	Development of tool for the assessment of comprehension of informed consent form in healthy volunteers participating in first-in-human studies
(Moseley, Wiggins et al. 2006)	Effects of presentation method on the understanding of informed consent
(Paris, Nogueira da Gama Chaves et al. 2007)	Improvement of the comprehension of written information given to healthy volunteers in biomedical research: a single-blind randomized controlled study
(Akkad, Jackson et al. 2006)	Patients' perceptions of written consent: questionnaire study
(Pilegaard and Ravn 2012)	Readability of patient information can be improved
(Pesudovs, Luscombe et al. 2006)	Recall from informed consent counselling for cataract surgery
(Oosthuizen, Burns et al. 2012)	The changing face of informed surgical consent
(Buckles, Powlishta et al. 2003)	Understanding of informed consent by demented individuals

Supplementary Table 14 Theme: Other

Author	Title
(Haux, Knaup et al. 2007)	On educating about medical data management the other side of the electronic health record
(Donnan, McLernon et al. 2009)	Development of a decision support tool for primary care management of patients with abnormal liver function tests without clinically apparent liver disease: a record-linkage population cohort study and decision analysis (ALFIE)
(Chulada, Vahdat et al. 2008)	The Environmental Polymorphisms Registry: a DNA resource to study genetic susceptibility loci
(Gracie, Lyon et al. 2010)	All Our Babies cohort study: recruitment of a cohort to predict women at risk of preterm birth through examination of gene expression profiles and the environment
(Ries, LeGrandeur et al. 2010)	Handling ethical, legal and social issues in birth cohort studies involving genetic research: responses from studies in six countries
(Hammerschmidt and Keane 1992)	Institutional review board (IRB) review lacks Impact on the readability of consent forms for research
(Kalton 2012)	Measuring health in population surveys
(Armstrong, Kline-Rogers et al. 2005)	Potential impact of the HIPAA privacy rule on data collection in a registry of patients with acute coronary syndrome
(Da Rocha and Seoane 2008)	Alternative consent models for biobanks: The new Spanish law on biomedical research
(Terry and Francis 2007)	Ensuring the privacy and confidentiality of electronic health records
(Ruan and Varadharajan 2003)	Supporting E-consent on Health data by Logic
(Kuchinke, Ohmann et al. 2014)	A standardised graphic method describing data privacy frameworks in primary care research using flexible zone model

Supplementary Figure 1 ALSPAC consent form

ALSPAC (http://www.bristol.ac.uk/alspac/)



University of BRISTOL

- Tick 'I AGREE' or 'I do NOT agree' for each section
- Sign, date and print your name
- Post this form back to us
- Sit back and relax!

CONSENT FORM Version 2 – 10/01/2011

Please do NOT fill this form in on behalf of another person.

If you have any questions, please contact us at: alspac-project@bristol.ac.uk or on 0117 33 10010.

1 Taking part in Children of the 90s

Children of the 90s (University of Bristol) would like to keep your contact details and a record of your involvement in the study. We will use these to invite you to take part in future Children of the 90s research.

I agree to Children of the 90s (University of Bristol) keeping a record of my contact details and a record of my involvement details for this purpose.

□ I AGREE
□ I do NOT agree

2 'Data Linkage' to access your official records

Children of the 90s would like to collect information about you that is stored in your official records. Your records will be used in Children of the 90s research. We will collect your records on a regular basis into the future, unless you tell us to stop. The records (and where they are held) are listed below.

Page 17 onwards of our information booklet explains 'data linkage', how we protect your confidentiality and how your records could help us with our research.

I understand that my details such as name, gender, date of birth, address, NHS number, education number and National Insurance number will be used to make an accurate link to my official records listed below.
(See Page 20 for more information)

Health Records (See page 23 for more information)

I agree to my healthcare provider and/or doctor (General Practitioner) being informed of my involvement in Children of the 90s.

and

I authorise my healthcare provider and/or doctor (General Practitioner) to provide my health records to Children of the 90s (University of Bristol).

and

I understand that information held by the NHS, National Health Service databases, and records maintained by The NHS Information Centre and the NHS Central Register may be used to help contact me and provide information about my health status.

□ I AGREE
□ I do NOT agree

Education Records (See page 24 for more information)

SCHOOL: I authorise the Department for Education to provide Children of the 90s (University of Bristol) with information from my education records for use in their research and/or for the purposes of contacting me.

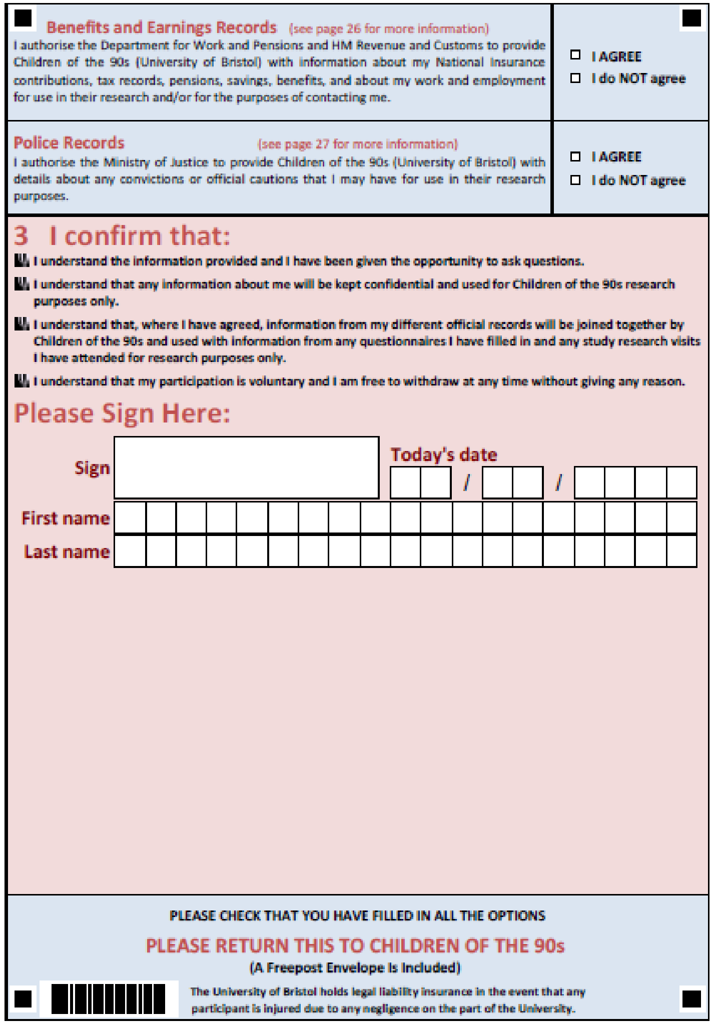
□ I AGREE
□ I do NOT agree

FURTHER EDUCATION: I authorise The Data Service and the Department for Business, Innovation & Skills to provide Children of the 90s (University of Bristol) with information from my education records for use in their research and/or for the purposes of contacting me.

□ I AGREE
□ I do NOT agree

HIGHER EDUCATION: I authorise the Universities and Colleges Admissions Service (UCAS) and the Higher Education Statistics Agency (HESA) to provide Children of the 90s (University of Bristol) with information from my education records for use in their research and/or for the purposes of contacting me.

□ I AGREE
□ I do NOT agree



Benefits and Earnings Records (see page 26 for more information)

I authorise the Department for Work and Pensions and HM Revenue and Customs to provide Children of the 90s (University of Bristol) with information about my National Insurance contributions, tax records, pensions, savings, benefits, and about my work and employment for use in their research and/or for the purposes of contacting me.

□ I AGREE
□ I do NOT agree

Police Records (see page 27 for more information)

I authorise the Ministry of Justice to provide Children of the 90s (University of Bristol) with details about any convictions or official cautions that I may have for use in their research purposes.

□ I AGREE
□ I do NOT agree

3 I confirm that:

- I understand the information provided and I have been given the opportunity to ask questions.
- I understand that any information about me will be kept confidential and used for Children of the 90s research purposes only.
- I understand that, where I have agreed, information from my different official records will be joined together by Children of the 90s and used with information from any questionnaires I have filled in and any study research visits I have attended for research purposes only.
- I understand that my participation is voluntary and I am free to withdraw at any time without giving any reason.

Please Sign Here:

Sign ______ Today's date __ __ / __ __ / __ __ __ __

First name ______

Last name ______

PLEASE CHECK THAT YOU HAVE FILLED IN ALL THE OPTIONS

PLEASE RETURN THIS TO CHILDREN OF THE 90s

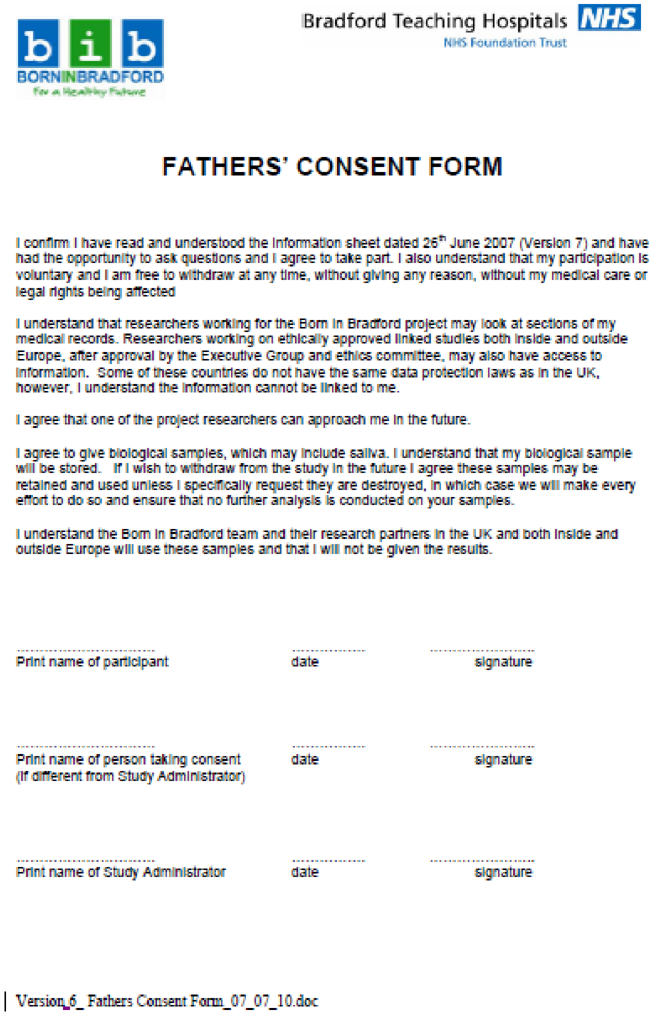
(A Freepost Envelope Is Included)

The University of Bristol holds legal liability insurance in the event that any participant is injured due to any negligence on the part of the University.

Supplementary Figure 2 Born in Bradford consent forms

Born in Bradford (https://borninbradford.nhs.uk/)

Father's consent form



BORNINBRADFORD
For a Healthy Future

Bradford Teaching Hospitals NHS
NHS Foundation Trust

FATHERS' CONSENT FORM

I confirm I have read and understood the information sheet dated 26th June 2007 (Version 7) and have had the opportunity to ask questions and I agree to take part. I also understand that my participation is voluntary and I am free to withdraw at any time, without giving any reason, without my medical care or legal rights being affected

I understand that researchers working for the Born in Bradford project may look at sections of my medical records. Researchers working on ethically approved linked studies both inside and outside Europe, after approval by the Executive Group and ethics committee, may also have access to information. Some of these countries do not have the same data protection laws as in the UK, however, I understand the information cannot be linked to me.

I agree that one of the project researchers can approach me in the future.

I agree to give biological samples, which may include saliva. I understand that my biological sample will be stored. If I wish to withdraw from the study in the future I agree these samples may be retained and used unless I specifically request they are destroyed, in which case we will make every effort to do so and ensure that no further analysis is conducted on your samples.

I understand the Born in Bradford team and their research partners in the UK and both inside and outside Europe will use these samples and that I will not be given the results.

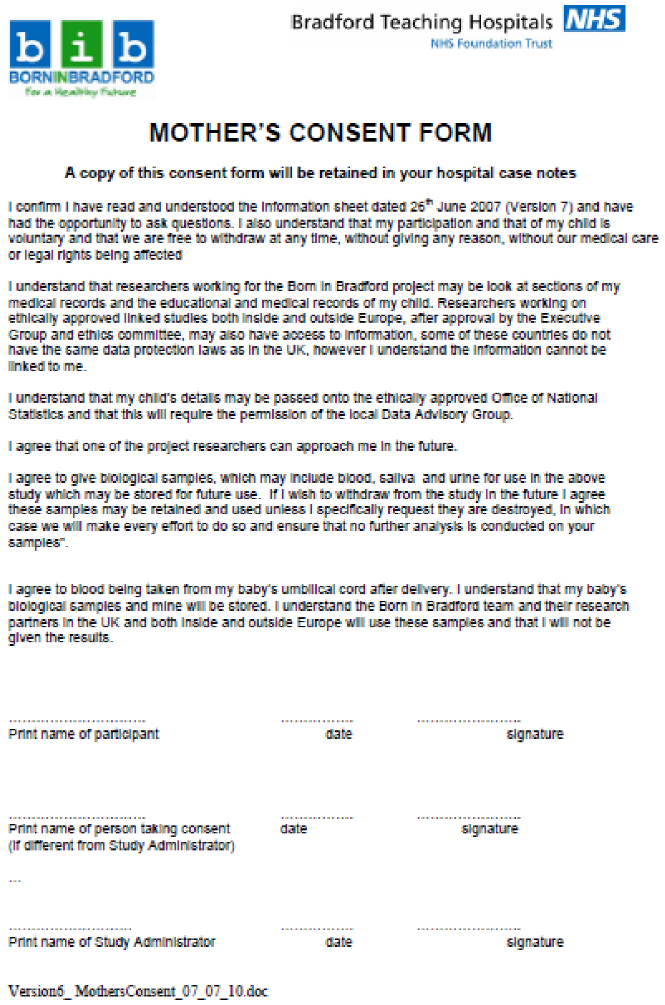
……………………………… ……………… ……………………
Print name of participant date signature

……………………………… ……………… ……………………
Print name of person taking consent date signature
(if different from Study Administrator)

……………………………… ……………… ……………………
Print name of Study Administrator date signature

Version_6_ Fathers Consent Form_07_07_10.doc

Mother's consent form



BORNINBRADFORD
For a Healthy Future

Bradford Teaching Hospitals NHS
NHS Foundation Trust

MOTHER'S CONSENT FORM

A copy of this consent form will be retained in your hospital case notes

I confirm I have read and understood the information sheet dated 26th June 2007 (Version 7) and have had the opportunity to ask questions. I also understand that my participation and that of my child is voluntary and that we are free to withdraw at any time, without giving any reason, without our medical care or legal rights being affected

I understand that researchers working for the Born in Bradford project may be look at sections of my medical records and the educational and medical records of my child. Researchers working on ethically approved linked studies both inside and outside Europe, after approval by the Executive Group and ethics committee, may also have access to information, some of these countries do not have the same data protection laws as in the UK, however I understand the information cannot be linked to me.

I understand that my child's details may be passed onto the ethically approved Office of National Statistics and that this will require the permission of the local Data Advisory Group.

I agree that one of the project researchers can approach me in the future.

I agree to give biological samples, which may include blood, saliva and urine for use in the above study which may be stored for future use. If I wish to withdraw from the study in the future I agree these samples may be retained and used unless I specifically request they are destroyed, in which case we will make every effort to do so and ensure that no further analysis is conducted on your samples".

I agree to blood being taken from my baby's umbilical cord after delivery. I understand that my baby's biological samples and mine will be stored. I understand the Born in Bradford team and their research partners in the UK and both inside and outside Europe will use these samples and that I will not be given the results.

……………………………… ……………… ……………………
Print name of participant date signature

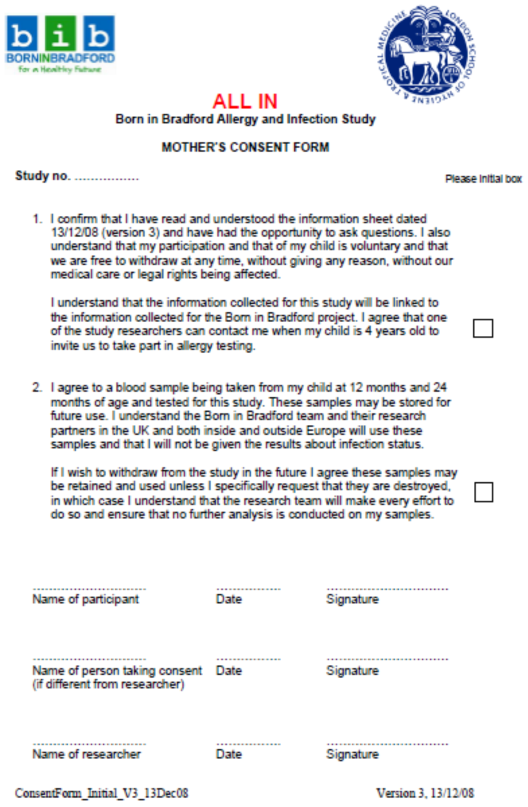
……………………………… ……………… ……………………
Print name of person taking consent date signature
(if different from Study Administrator)

…

……………………………… ……………… ……………………
Print name of Study Administrator date signature

Version6_ MothersConsent_07_07_10.doc

Allergy and Infection Study



ALL IN

Born in Bradford Allergy and Infection Study

MOTHER'S CONSENT FORM

Study no. .... Please initial box

1. I confirm that I have read and understood the information sheet dated 13/12/08 (version 3) and have had the opportunity to ask questions. I also understand that my participation and that of my child is voluntary and that we are free to withdraw at any time, without giving any reason, without our medical care or legal rights being affected.

   I understand that the information collected for this study will be linked to the information collected for the Born in Bradford project. I agree that one of the study researchers can contact me when my child is 4 years old to invite us to take part in allergy testing. □

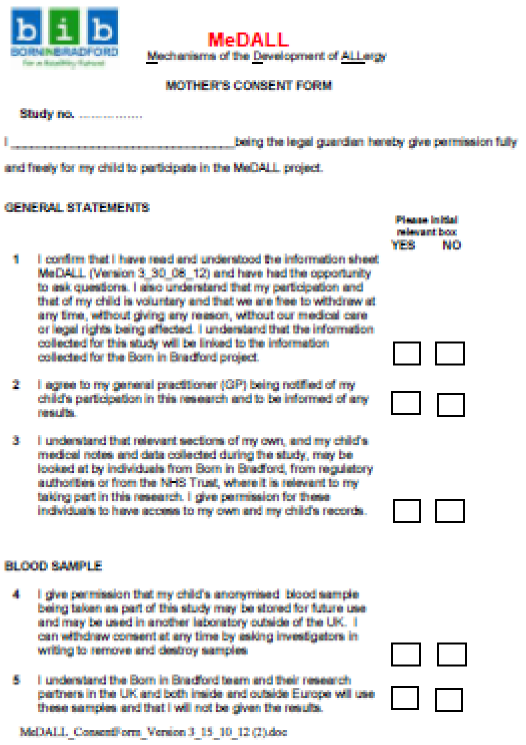
2. I agree to a blood sample being taken from my child at 12 months and 24 months of age and tested for this study. These samples may be stored for future use. I understand the Born in Bradford team and their research partners in the UK and both inside and outside Europe will use these samples and that I will not be given the results about infection status.

   If I wish to withdraw from the study in the future I agree these samples may be retained and used unless I specifically request that they are destroyed, in which case I understand that the research team will make every effort to do so and ensure that no further analysis is conducted on my samples. □

.... Name of participant	.... Date	.... Signature
.... Name of person taking consent (if different from researcher)	.... Date	.... Signature
.... Name of researcher	.... Date	.... Signature

ConsentForm_Initial_V3_13Dec08 Version 3, 13/12/08

Mechanisms of the Development of Allergy



MeDALL

Mechanisms of the Development of ALLergy

MOTHER'S CONSENT FORM

Study no. ....

I _________________________________being the legal guardian hereby give permission fully and freely for my child to participate in the MeDALL project.

GENERAL STATEMENTS

Please initial relevant box
YES NO

1 I confirm that I have read and understood the information sheet MeDALL (Version 3_30_08_12) and have had the opportunity to ask questions. I also understand that my participation and that of my child is voluntary and that we are free to withdraw at any time, without giving any reason, without our medical care or legal rights being affected. I understand that the information collected for this study will be linked to the information collected for the Born in Bradford project. □ □

2 I agree to my general practitioner (GP) being notified of my child's participation in this research and to be informed of any results. □ □

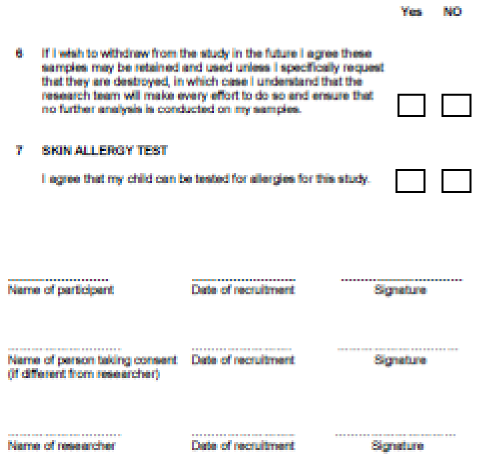
3 I understand that relevant sections of my own, and my child's medical notes and data collected during the study, may be looked at by individuals from Born in Bradford, from regulatory authorities or from the NHS Trust, where it is relevant to my taking part in this research. I give permission for these individuals to have access to my own and my child's records. □ □

BLOOD SAMPLE

4 I give permission that my child's anonymised blood sample being taken as part of this study may be stored for future use and may be used in another laboratory outside of the UK. I can withdraw consent at any time by asking investigators in writing to remove and destroy samples □ □

5 I understand the Born in Bradford team and their research partners in the UK and both inside and outside Europe will use these samples and that I will not be given the results. □ □

MeDALL_ConsentForm_Version 3_15_10_12 (2).doc



Yes NO

6 If I wish to withdraw from the study in the future I agree these samples may be retained and used unless I specifically request that they are destroyed, in which case I understand that the research team will make every effort to do so and ensure that no further analysis is conducted on my samples. □ □

7 **SKIN ALLERGY TEST**

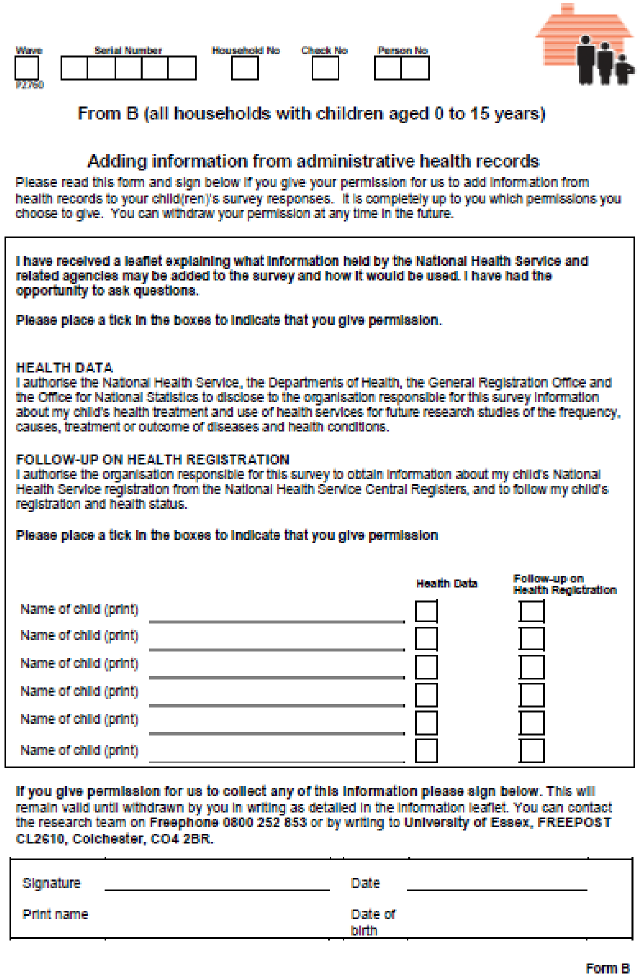
I agree that my child can be tested for allergies for this study. □ □

.... Name of participant	.... Date of recruitment	.... Signature
.... Name of person taking consent (if different from researcher)	.... Date of recruitment	.... Signature
.... Name of researcher	.... Date of recruitment	.... Signature

Supplementary Figure 3 British Household Panel Survey consent forms

British Household Panel Survey (https://www.iser.essex.ac.uk/bhps/)

Form B



Wave	Serial Number	Household No	Check No	Person No
□	□□□□□	□	□	□□

P2760

From B (all households with children aged 0 to 15 years)

Adding information from administrative health records

Please read this form and sign below if you give your permission for us to add information from health records to your child(ren)'s survey responses. It is completely up to you which permissions you choose to give. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what information held by the National Health Service and related agencies may be added to the survey and how it would be used. I have had the opportunity to ask questions.

Please place a tick in the boxes to indicate that you give permission.

HEALTH DATA
I authorise the National Health Service, the Departments of Health, the General Registration Office and the Office for National Statistics to disclose to the organisation responsible for this survey information about my child's health treatment and use of health services for future research studies of the frequency, causes, treatment or outcome of diseases and health conditions.

FOLLOW-UP ON HEALTH REGISTRATION
I authorise the organisation responsible for this survey to obtain information about my child's National Health Service registration from the National Health Service Central Registers, and to follow my child's registration and health status.

Please place a tick in the boxes to indicate that you give permission

	Health Data	Follow-up on Health Registration
Name of child (print) ____________	□	□
Name of child (print) ____________	□	□
Name of child (print) ____________	□	□
Name of child (print) ____________	□	□
Name of child (print) ____________	□	□
Name of child (print) ____________	□	□

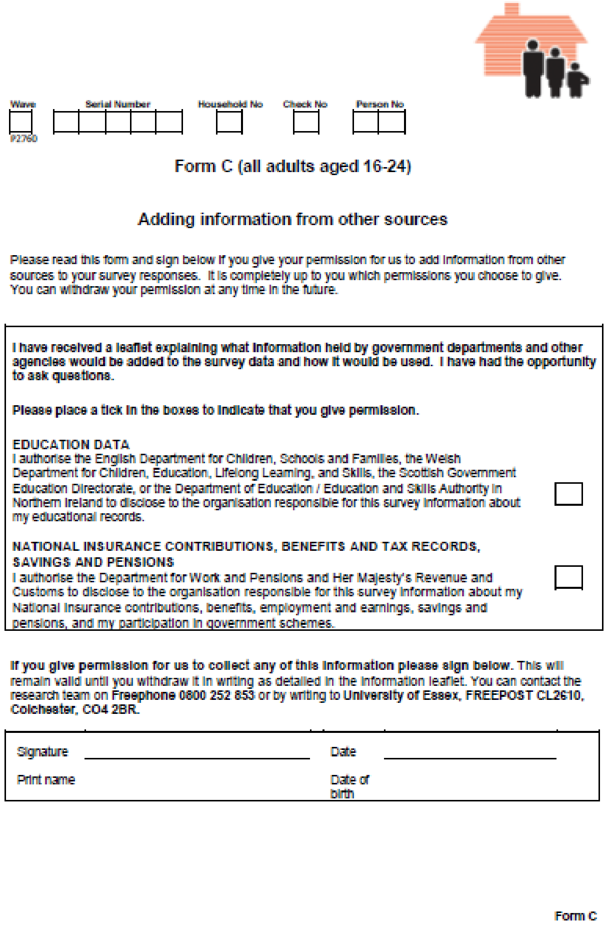
If you give permission for us to collect any of this information please sign below. This will remain valid until withdrawn by you in writing as detailed in the information leaflet. You can contact the research team on **Freephone 0800 252 853** or by writing to **University of Essex, FREEPOST CL2610, Colchester, CO4 2BR.**

Signature ____________ Date ____________

Print name Date of birth

Form B

Form C



Wave	Serial Number	Household No	Check No	Person No
□	□□□□□	□	□	□□

P2760

Form C (all adults aged 16-24)

Adding information from other sources

Please read this form and sign below if you give your permission for us to add information from other sources to your survey responses. It is completely up to you which permissions you choose to give. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what information held by government departments and other agencies would be added to the survey data and how it would be used. I have had the opportunity to ask questions.

Please place a tick in the boxes to indicate that you give permission.

EDUCATION DATA
I authorise the English Department for Children, Schools and Families, the Welsh Department for Children, Education, Lifelong Learning, and Skills, the Scottish Government Education Directorate, or the Department of Education / Education and Skills Authority in Northern Ireland to disclose to the organisation responsible for this survey information about my educational records. □

NATIONAL INSURANCE CONTRIBUTIONS, BENEFITS AND TAX RECORDS, SAVINGS AND PENSIONS
I authorise the Department for Work and Pensions and Her Majesty's Revenue and Customs to disclose to the organisation responsible for this survey information about my National Insurance contributions, benefits, employment and earnings, savings and pensions, and my participation in government schemes. □

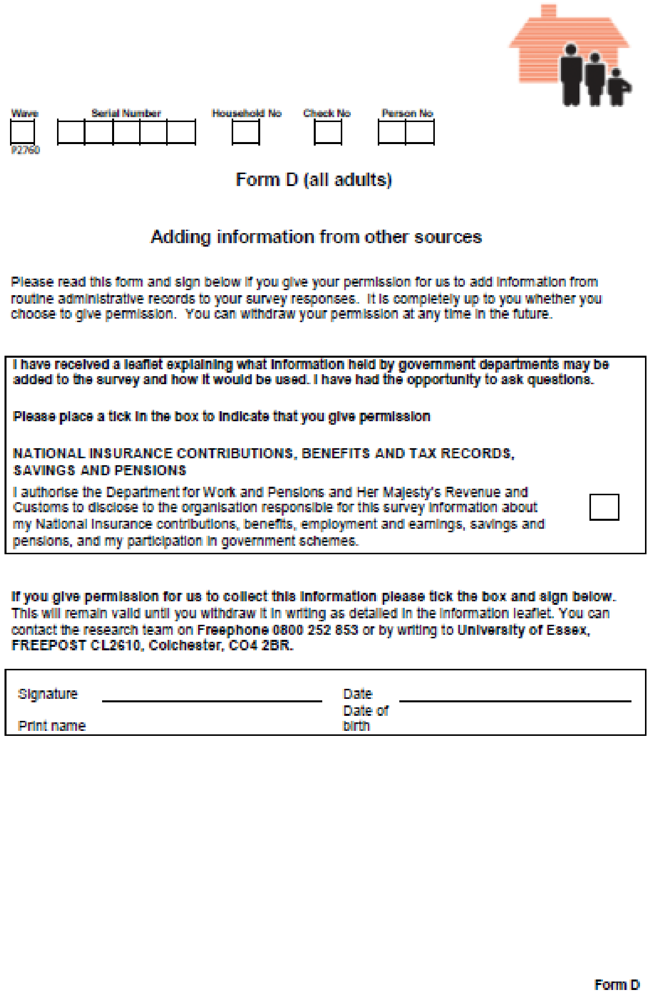
If you give permission for us to collect any of this information please sign below. This will remain valid until you withdraw it in writing as detailed in the information leaflet. You can contact the research team on **Freephone 0800 252 853** or by writing to **University of Essex, FREEPOST CL2610, Colchester, CO4 2BR.**

Signature ____________ Date ____________

Print name Date of birth

Form C

Form D



Wave □ P2760 Serial Number □□□□□ Household No □ Check No □ Person No □□

Form D (all adults)

Adding information from other sources

Please read this form and sign below if you give your permission for us to add information from routine administrative records to your survey responses. It is completely up to you whether you choose to give permission. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what information held by government departments may be added to the survey and how it would be used. I have had the opportunity to ask questions.

Please place a tick in the box to indicate that you give permission

NATIONAL INSURANCE CONTRIBUTIONS, BENEFITS AND TAX RECORDS, SAVINGS AND PENSIONS

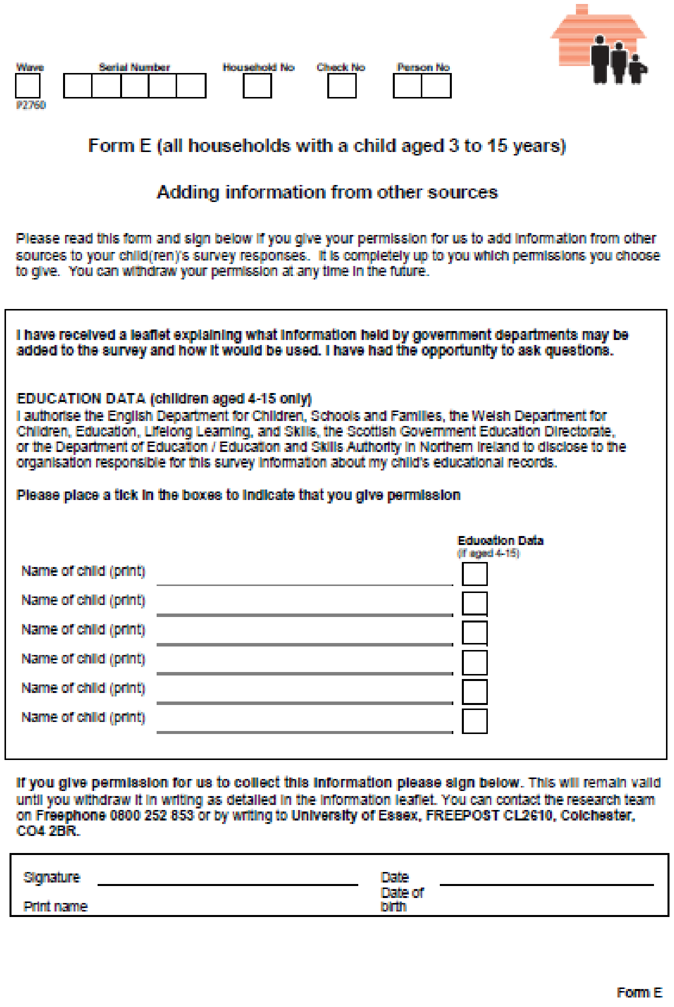
I authorise the Department for Work and Pensions and Her Majesty's Revenue and Customs to disclose to the organisation responsible for this survey information about my National Insurance contributions, benefits, employment and earnings, savings and pensions, and my participation in government schemes. □

If you give permission for us to collect this information please tick the box and sign below. This will remain valid until you withdraw it in writing as detailed in the information leaflet. You can contact the research team on **Freephone 0800 252 853** or by writing to **University of Essex, FREEPOST CL2610, Colchester, CO4 2BR.**

Signature ____________ Date ____________

Print name Date of birth

Form E



Wave □ P2760 Serial Number □□□□□ Household No □ Check No □ Person No □□

Form E (all households with a child aged 3 to 15 years)

Adding information from other sources

Please read this form and sign below if you give your permission for us to add information from other sources to your child(ren)'s survey responses. It is completely up to you which permissions you choose to give. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what information held by government departments may be added to the survey and how it would be used. I have had the opportunity to ask questions.

EDUCATION DATA (children aged 4-15 only)

I authorise the English Department for Children, Schools and Families, the Welsh Department for Children, Education, Lifelong Learning, and Skills, the Scottish Government Education Directorate, or the Department of Education / Education and Skills Authority in Northern Ireland to disclose to the organisation responsible for this survey information about my child's educational records.

Please place a tick in the boxes to indicate that you give permission

		Education Data (if aged 4-15)
Name of child (print)	____________	□
Name of child (print)	____________	□
Name of child (print)	____________	□
Name of child (print)	____________	□
Name of child (print)	____________	□
Name of child (print)	____________	□

If you give permission for us to collect this information please sign below. This will remain valid until you withdraw it in writing as detailed in the information leaflet. You can contact the research team on **Freephone 0800 252 853** or by writing to **University of Essex, FREEPOST CL2610, Colchester, CO4 2BR.**

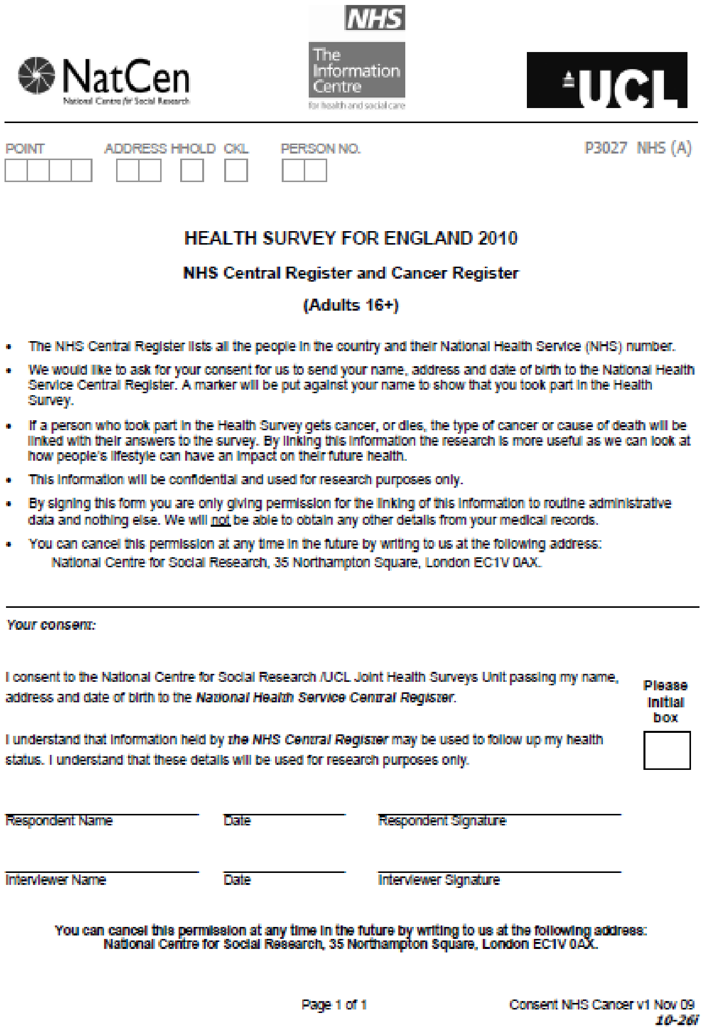
Signature ____________ Date ____________

Print name Date of birth

Supplementary Figure 4 Health Survey for England consent forms

Health Survey for England (Mindell, J., P. Biddulph, et al. (2012). “Cohort Profile: The Health Survey for England.” Int J Epidemiol **41**(6): 1585-1593)

NHS Central Register and Cancer Register – Adults 16+



NatCen National Centre for Social Research

NHS The Information Centre for health and social care

UCL

POINT □□□□ ADDRESS □□ HHOLD □ CKL □ PERSON NO. □□

P3027 NHS (A)

HEALTH SURVEY FOR ENGLAND 2010

NHS Central Register and Cancer Register

(Adults 16+)

- The NHS Central Register lists all the people in the country and their National Health Service (NHS) number.
- We would like to ask for your consent for us to send your name, address and date of birth to the National Health Service Central Register. A marker will be put against your name to show that you took part in the Health Survey.
- If a person who took part in the Health Survey gets cancer, or dies, the type of cancer or cause of death will be linked with their answers to the survey. By linking this information the research is more useful as we can look at how people's lifestyle can have an impact on their future health.
- This information will be confidential and used for research purposes only.
- By signing this form you are only giving permission for the linking of this information to routine administrative data and nothing else. We will not be able to obtain any other details from your medical records.
- You can cancel this permission at any time in the future by writing to us at the following address:
  National Centre for Social Research, 35 Northampton Square, London EC1V 0AX.

Your consent:

I consent to the National Centre for Social Research /UCL Joint Health Surveys Unit passing my name, address and date of birth to the *National Health Service Central Register*.

I understand that information held by *the NHS Central Register* may be used to follow up my health status. I understand that these details will be used for research purposes only.

Please initial box □

Respondent Name ____ Date ____ Respondent Signature ____

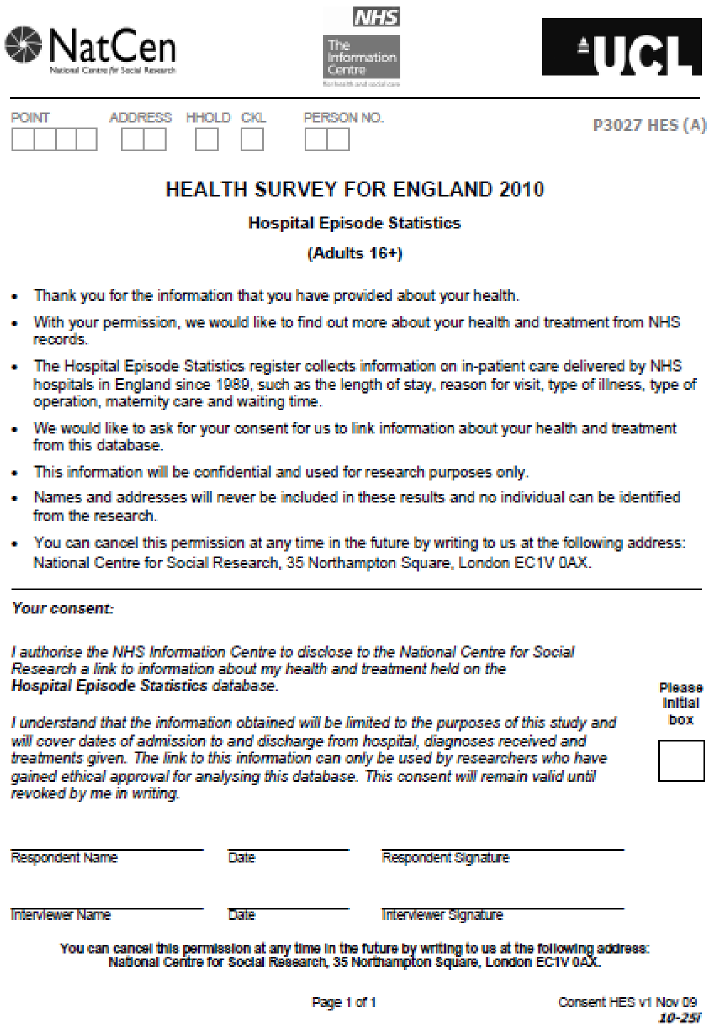
Interviewer Name ____ Date ____ Interviewer Signature ____

You can cancel this permission at any time in the future by writing to us at the following address:
National Centre for Social Research, 35 Northampton Square, London EC1V 0AX.

Page 1 of 1

Consent NHS Cancer v1 Nov 09
10-26i

Hospital Episode Statistics – Adults 16+



NatCen National Centre for Social Research

NHS The Information Centre for health and social care

UCL

POINT □□□□ ADDRESS □□ HHOLD □ CKL □ PERSON NO. □□

P3027 HES (A)

HEALTH SURVEY FOR ENGLAND 2010

Hospital Episode Statistics

(Adults 16+)

- Thank you for the information that you have provided about your health.
- With your permission, we would like to find out more about your health and treatment from NHS records.
- The Hospital Episode Statistics register collects information on in-patient care delivered by NHS hospitals in England since 1989, such as the length of stay, reason for visit, type of illness, type of operation, maternity care and waiting time.
- We would like to ask for your consent for us to link information about your health and treatment from this database.
- This information will be confidential and used for research purposes only.
- Names and addresses will never be included in these results and no individual can be identified from the research.
- You can cancel this permission at any time in the future by writing to us at the following address: National Centre for Social Research, 35 Northampton Square, London EC1V 0AX.

Your consent:

I authorise the NHS Information Centre to disclose to the National Centre for Social Research a link to information about my health and treatment held on the ***Hospital Episode Statistics*** *database.*

I understand that the information obtained will be limited to the purposes of this study and will cover dates of admission to and discharge from hospital, diagnoses received and treatments given. The link to this information can only be used by researchers who have gained ethical approval for analysing this database. This consent will remain valid until revoked by me in writing.

Please initial box □

Respondent Name ____ Date ____ Respondent Signature ____

Interviewer Name ____ Date ____ Interviewer Signature ____

You can cancel this permission at any time in the future by writing to us at the following address:
National Centre for Social Research, 35 Northampton Square, London EC1V 0AX.

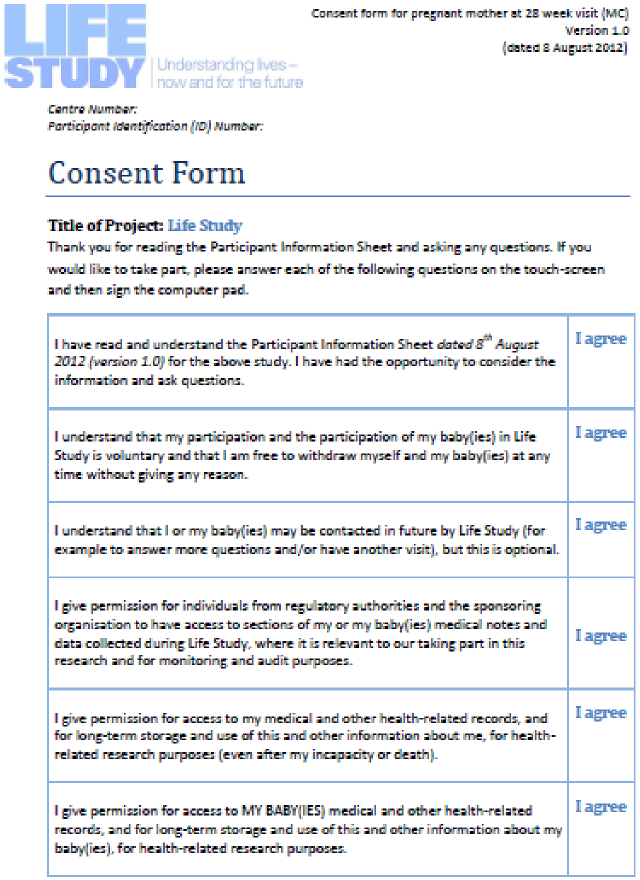
Page 1 of 1

Consent HES v1 Nov 09
10-25i

Supplementary Figure 5 Life Study consent forms

Life Study (http://www.lifestudy.ac.uk/)

Consent form for pregnant mother at 28 weeks



Understanding lives – now and for the future

Consent form for pregnant mother at 28 week visit (MC)
Version 1.0
(dated 8 August 2012)

Centre Number:
Participant Identification (ID) Number:

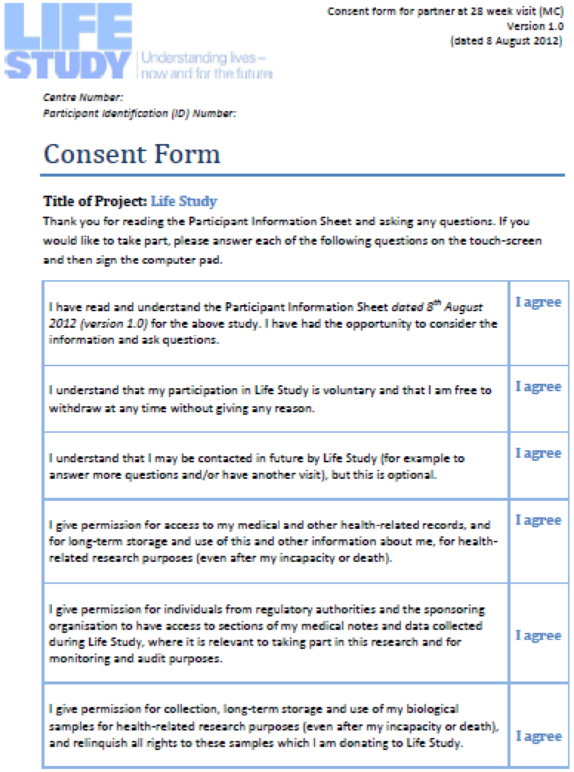
Consent Form

Title of Project: Life Study

Thank you for reading the Participant Information Sheet and asking any questions. If you would like to take part, please answer each of the following questions on the touch-screen and then sign the computer pad.

I have read and understand the Participant Information Sheet *dated 8th August 2012 (version 1.0)* for the above study. I have had the opportunity to consider the information and ask questions.	I agree
I understand that my participation and the participation of my baby(ies) in Life Study is voluntary and that I am free to withdraw myself and my baby(ies) at any time without giving any reason.	I agree
I understand that I or my baby(ies) may be contacted in future by Life Study (for example to answer more questions and/or have another visit), but this is optional.	I agree
I give permission for individuals from regulatory authorities and the sponsoring organisation to have access to sections of my or my baby(ies) medical notes and data collected during Life Study, where it is relevant to our taking part in this research and for monitoring and audit purposes.	I agree
I give permission for access to my medical and other health-related records, and for long-term storage and use of this and other information about me, for health-related research purposes (even after my incapacity or death).	I agree
I give permission for access to MY BABY(IES) medical and other health-related records, and for long-term storage and use of this and other information about my baby(ies), for health-related research purposes.	I agree

Consent form for partner 28 weeks



Understanding lives – now and for the future

Consent form for partner at 28 week visit (MC)
Version 1.0
(dated 8 August 2012)

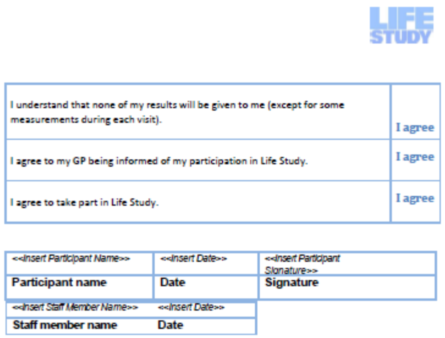
Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

Thank you for reading the Participant Information Sheet and asking any questions. If you would like to take part, please answer each of the following questions on the touch-screen and then sign the computer pad.

I have read and understand the Participant Information Sheet *dated 8th August 2012 (version 1.0)* for the above study. I have had the opportunity to consider the information and ask questions.	I agree
I understand that my participation in Life Study is voluntary and that I am free to withdraw at any time without giving any reason.	I agree
I understand that I may be contacted in future by Life Study (for example to answer more questions and/or have another visit), but this is optional.	I agree
I give permission for access to my medical and other health-related records, and for long-term storage and use of this and other information about me, for health-related research purposes (even after my incapacity or death).	I agree
I give permission for individuals from regulatory authorities and the sponsoring organisation to have access to sections of my medical notes and data collected during Life Study, where it is relevant to taking part in this research and for monitoring and audit purposes.	I agree
I give permission for collection, long-term storage and use of my biological samples for health-related research purposes (even after my incapacity or death), and relinquish all rights to these samples which I am donating to Life Study.	I agree

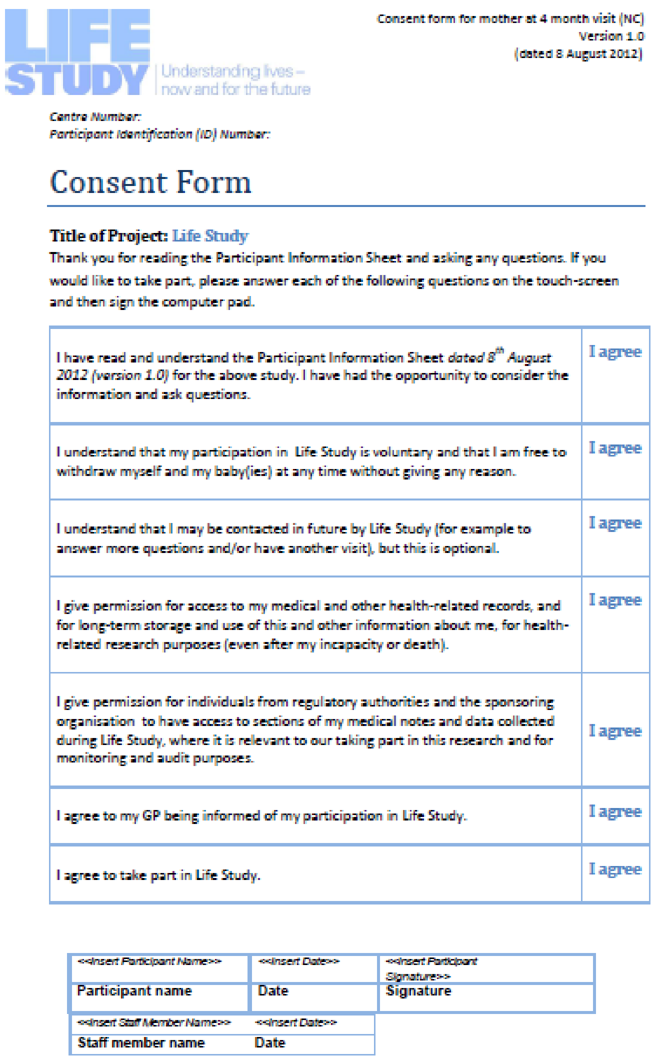


I understand that none of my results will be given to me (except for some measurements during each visit).	I agree
I agree to my GP being informed of my participation in Life Study.	I agree
I agree to take part in Life Study.	I agree

<<Insert Participant Name>>	*<<Insert Date>>*	*<<Insert Participant Signature>>*
Participant name	**Date**	**Signature**

<<Insert Staff Member Name>>	*<<Insert Date>>*
Staff member name	**Date**

Consent form for mother at 4 month visit



LIFE STUDY Understanding lives – now and for the future

Consent form for mother at 4 month visit (NC)
Version 1.0
(dated 8 August 2012)

Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

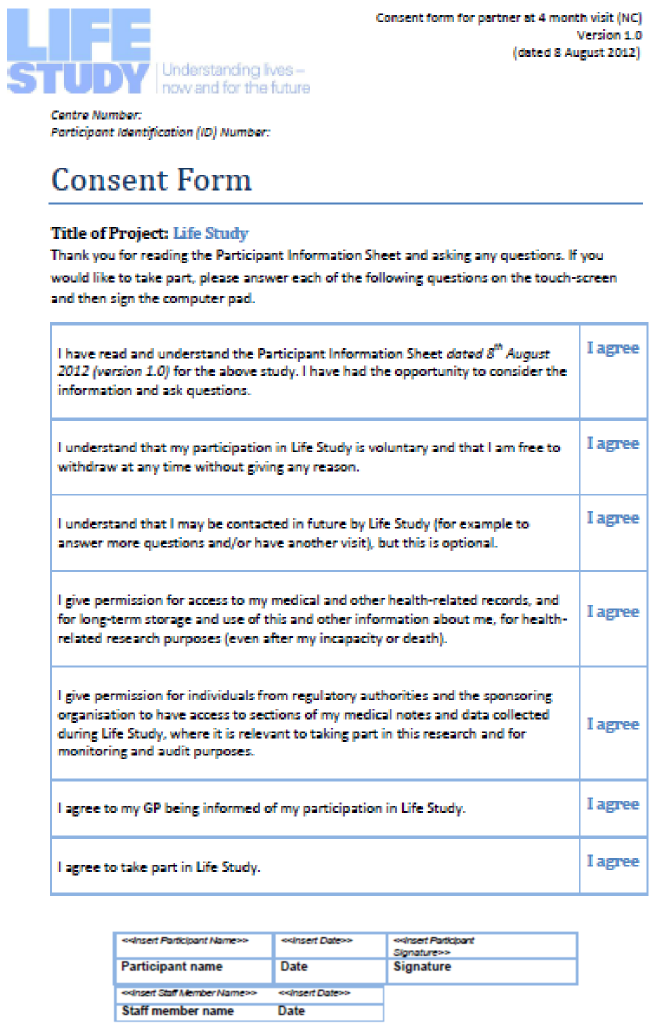
Thank you for reading the Participant Information Sheet and asking any questions. If you would like to take part, please answer each of the following questions on the touch-screen and then sign the computer pad.

I have read and understand the Participant Information Sheet *dated 8th August 2012 (version 1.0)* for the above study. I have had the opportunity to consider the information and ask questions.	I agree
I understand that my participation in Life Study is voluntary and that I am free to withdraw myself and my baby(ies) at any time without giving any reason.	I agree
I understand that I may be contacted in future by Life Study (for example to answer more questions and/or have another visit), but this is optional.	I agree
I give permission for access to my medical and other health-related records, and for long-term storage and use of this and other information about me, for health-related research purposes (even after my incapacity or death).	I agree
I give permission for individuals from regulatory authorities and the sponsoring organisation to have access to sections of my medical notes and data collected during Life Study, where it is relevant to our taking part in this research and for monitoring and audit purposes.	I agree
I agree to my GP being informed of my participation in Life Study.	I agree
I agree to take part in Life Study.	I agree

<<Insert Participant Name>>	*<<Insert Date>>*	*<<Insert Participant Signature>>*
Participant name	**Date**	**Signature**

<<Insert Staff Member Name>>	*<<Insert Date>>*
Staff member name	**Date**

Consent form for partner 4 visit



LIFE STUDY Understanding lives – now and for the future

Consent form for partner at 4 month visit (NC)
Version 1.0
(dated 8 August 2012)

Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

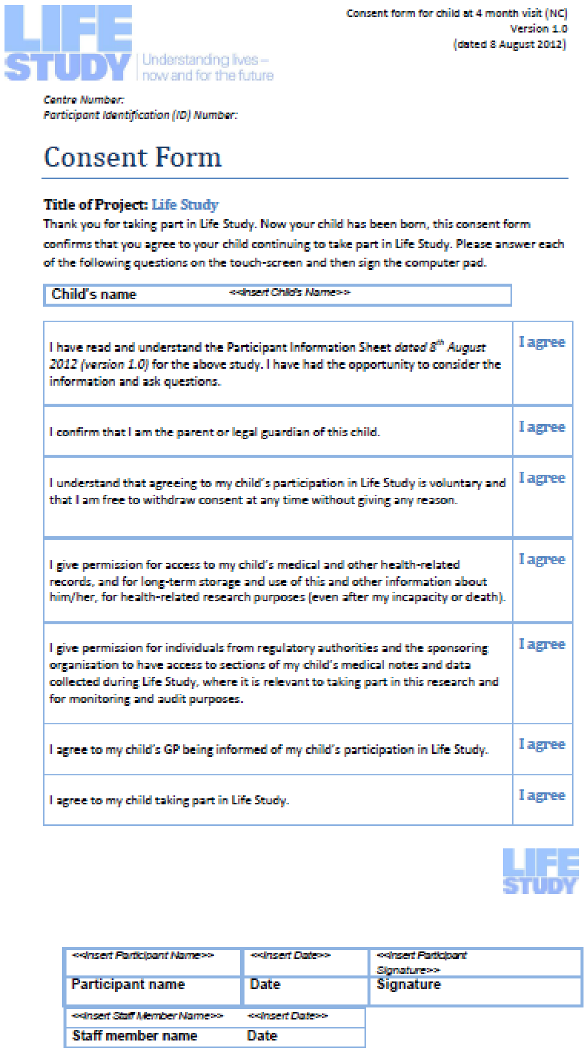
Thank you for reading the Participant Information Sheet and asking any questions. If you would like to take part, please answer each of the following questions on the touch-screen and then sign the computer pad.

I have read and understand the Participant Information Sheet *dated 8th August 2012 (version 1.0)* for the above study. I have had the opportunity to consider the information and ask questions.	I agree
I understand that my participation in Life Study is voluntary and that I am free to withdraw at any time without giving any reason.	I agree
I understand that I may be contacted in future by Life Study (for example to answer more questions and/or have another visit), but this is optional.	I agree
I give permission for access to my medical and other health-related records, and for long-term storage and use of this and other information about me, for health-related research purposes (even after my incapacity or death).	I agree
I give permission for individuals from regulatory authorities and the sponsoring organisation to have access to sections of my medical notes and data collected during Life Study, where it is relevant to taking part in this research and for monitoring and audit purposes.	I agree
I agree to my GP being informed of my participation in Life Study.	I agree
I agree to take part in Life Study.	I agree

<<Insert Participant Name>>	*<<Insert Date>>*	*<<Insert Participant Signature>>*
Participant name	**Date**	**Signature**

<<Insert Staff Member Name>>	*<<Insert Date>>*
Staff member name	**Date**

Consent form for child at 4 month visit



LIFE STUDY Understanding lives – now and for the future

Consent form for child at 4 month visit (NC)
Version 1.0
(dated 8 August 2012)

Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

Thank you for taking part in Life Study. Now your child has been born, this consent form confirms that you agree to your child continuing to take part in Life Study. Please answer each of the following questions on the touch-screen and then sign the computer pad.

Child's name	*<<Insert Child's Name>>*

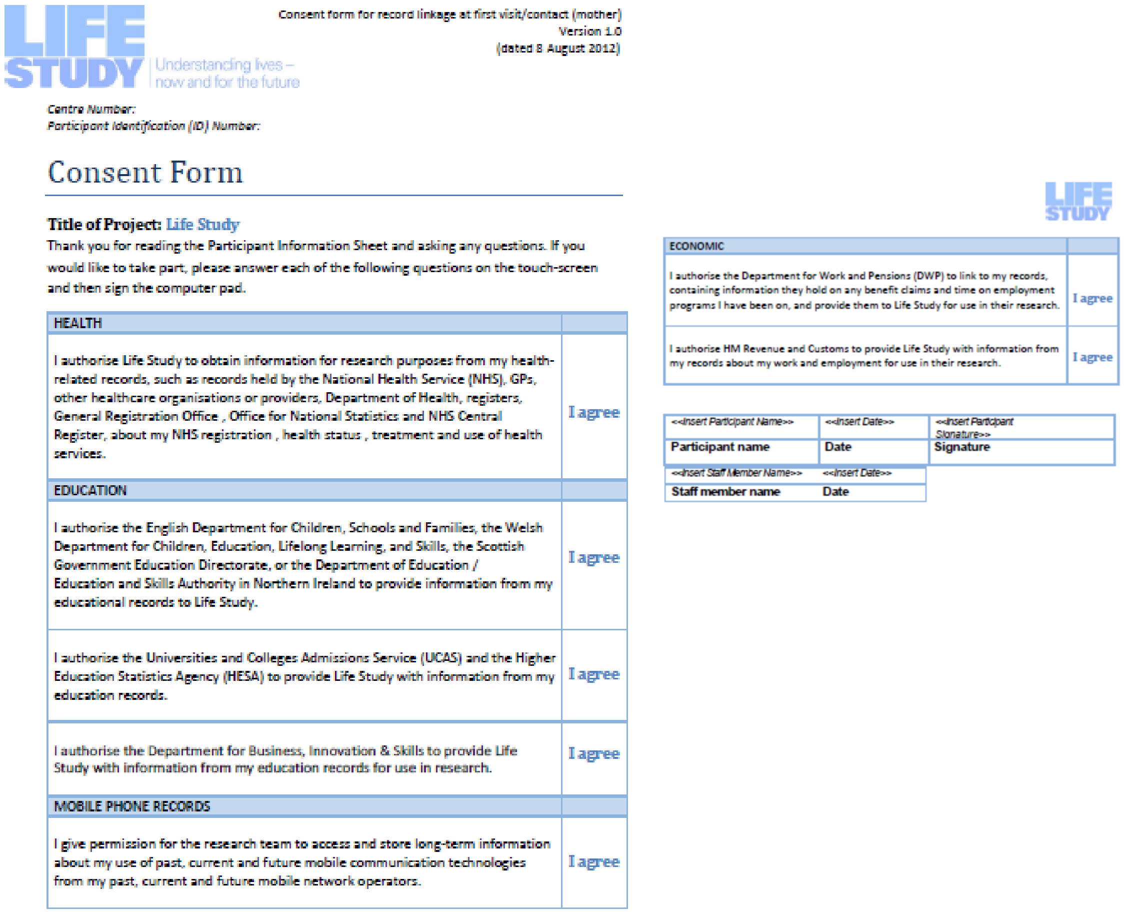
I have read and understand the Participant Information Sheet *dated 8th August 2012 (version 1.0)* for the above study. I have had the opportunity to consider the information and ask questions.	I agree
I confirm that I am the parent or legal guardian of this child.	I agree
I understand that agreeing to my child's participation in Life Study is voluntary and that I am free to withdraw consent at any time without giving any reason.	I agree
I give permission for access to my child's medical and other health-related records, and for long-term storage and use of this and other information about him/her, for health-related research purposes (even after my incapacity or death).	I agree
I give permission for individuals from regulatory authorities and the sponsoring organisation to have access to sections of my child's medical notes and data collected during Life Study, where it is relevant to taking part in this research and for monitoring and audit purposes.	I agree
I agree to my child's GP being informed of my child's participation in Life Study.	I agree
I agree to my child taking part in Life Study.	I agree

LIFE STUDY

<<Insert Participant Name>>	*<<Insert Date>>*	*<<Insert Participant Signature>>*
Participant name	**Date**	**Signature**

<<Insert Staff Member Name>>	*<<Insert Date>>*
Staff member name	**Date**

Consent form for record linkage at first visit/contact (mother)



LIFE STUDY Understanding lives – now and for the future

Consent form for record linkage at first visit/contact (mother)
Version 1.0
(dated 8 August 2012)

Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

Thank you for reading the Participant Information Sheet and asking any questions. If you would like to take part, please answer each of the following questions on the touch-screen and then sign the computer pad.

HEALTH	
I authorise Life Study to obtain information for research purposes from my health-related records, such as records held by the National Health Service (NHS), GPs, other healthcare organisations or providers, Department of Health, registers, General Registration Office , Office for National Statistics and NHS Central Register, about my NHS registration , health status , treatment and use of health services.	I agree
EDUCATION	
I authorise the English Department for Children, Schools and Families, the Welsh Department for Children, Education, Lifelong Learning, and Skills, the Scottish Government Education Directorate, or the Department of Education / Education and Skills Authority in Northern Ireland to provide information from my educational records to Life Study.	I agree
I authorise the Universities and Colleges Admissions Service (UCAS) and the Higher Education Statistics Agency (HESA) to provide Life Study with information from my education records.	I agree
I authorise the Department for Business, Innovation & Skills to provide Life Study with information from my education records for use in research.	I agree
MOBILE PHONE RECORDS	
I give permission for the research team to access and store long-term information about my use of past, current and future mobile communication technologies from my past, current and future mobile network operators.	I agree

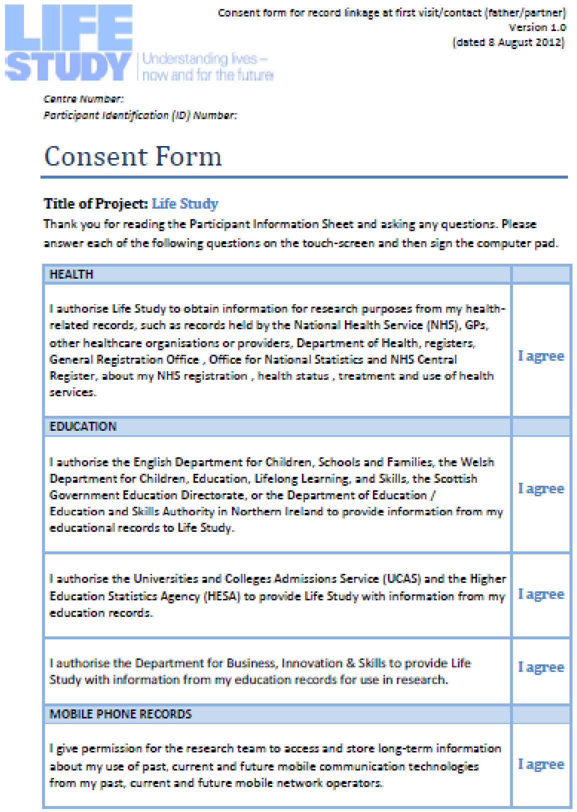
LIFE STUDY

ECONOMIC	
I authorise the Department for Work and Pensions (DWP) to link to my records, containing information they hold on any benefit claims and time on employment programs I have been on, and provide them to Life Study for use in their research.	I agree
I authorise HM Revenue and Customs to provide Life Study with information from my records about my work and employment for use in their research.	I agree

<<Insert Participant Name>>	*<<Insert Date>>*	*<<Insert Participant Signature>>*
Participant name	**Date**	**Signature**

<<Insert Staff Member Name>>	*<<Insert Date>>*
Staff member name	**Date**

Consent form for record linkage at first visit/contact (father/partner)



LIFE STUDY | Understanding lives – now and for the future

Consent form for record linkage at first visit/contact (father/partner)
Version 1.0
(dated 8 August 2012)

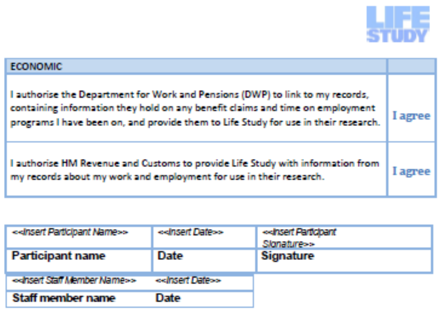
Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

Thank you for reading the Participant Information Sheet and asking any questions. Please answer each of the following questions on the touch-screen and then sign the computer pad.

HEALTH	
I authorise Life Study to obtain information for research purposes from my health-related records, such as records held by the National Health Service (NHS), GPs, other healthcare organisations or providers, Department of Health, registers, General Registration Office , Office for National Statistics and NHS Central Register, about my NHS registration , health status , treatment and use of health services.	I agree
EDUCATION	
I authorise the English Department for Children, Schools and Families, the Welsh Department for Children, Education, Lifelong Learning, and Skills, the Scottish Government Education Directorate, or the Department of Education / Education and Skills Authority in Northern Ireland to provide information from my educational records to Life Study.	I agree
I authorise the Universities and Colleges Admissions Service (UCAS) and the Higher Education Statistics Agency (HESA) to provide Life Study with information from my education records.	I agree
I authorise the Department for Business, Innovation & Skills to provide Life Study with information from my education records for use in research.	I agree
MOBILE PHONE RECORDS	
I give permission for the research team to access and store long-term information about my use of past, current and future mobile communication technologies from my past, current and future mobile network operators.	I agree



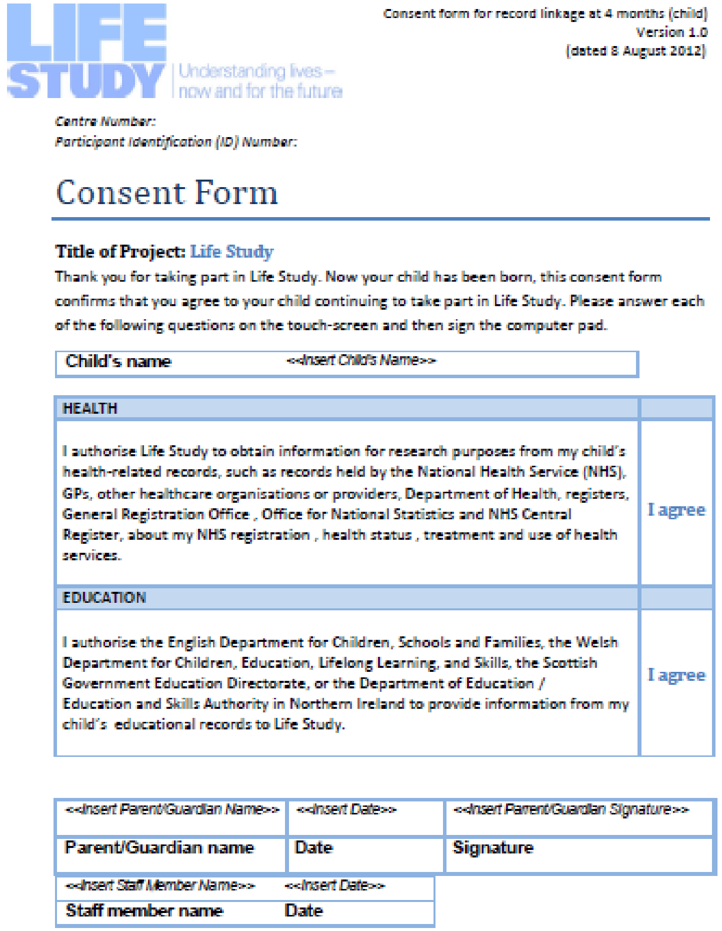
LIFE STUDY

ECONOMIC	
I authorise the Department for Work and Pensions (DWP) to link to my records, containing information they hold on any benefit claims and time on employment programs I have been on, and provide them to Life Study for use in their research.	I agree
I authorise HM Revenue and Customs to provide Life Study with information from my records about my work and employment for use in their research.	I agree

<<Insert Participant Name>>	*<<Insert Date>>*	*<<Insert Participant Signature>>*
Participant name	**Date**	**Signature**

<<Insert Staff Member Name>>	*<<Insert Date>>*
Staff member name	**Date**

Consent form for record linkage at 4 months (child)



LIFE STUDY | Understanding lives – now and for the future

Consent form for record linkage at 4 months (child)
Version 1.0
(dated 8 August 2012)

Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

Thank you for taking part in Life Study. Now your child has been born, this consent form confirms that you agree to your child continuing to take part in Life Study. Please answer each of the following questions on the touch-screen and then sign the computer pad.

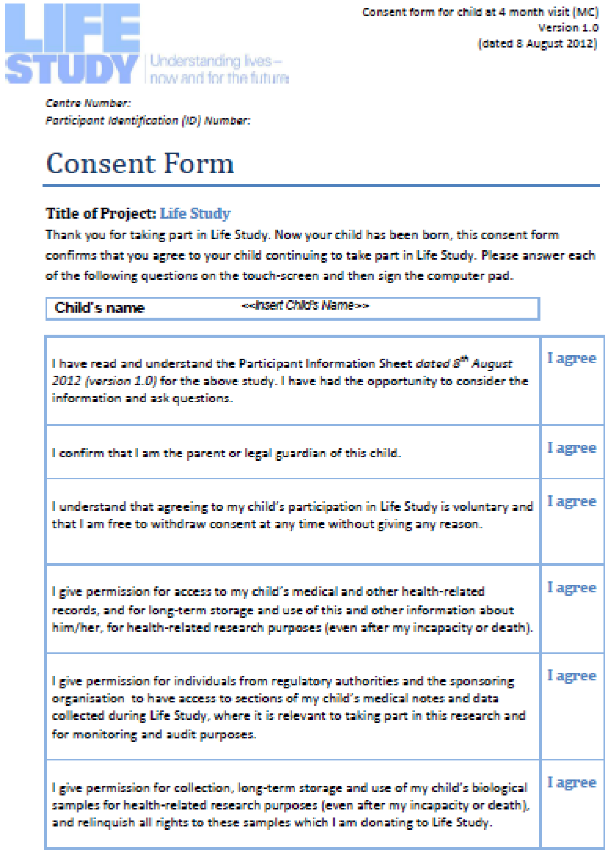
Child's name	*<<Insert Child's Name>>*

HEALTH	
I authorise Life Study to obtain information for research purposes from my child's health-related records, such as records held by the National Health Service (NHS), GPs, other healthcare organisations or providers, Department of Health, registers, General Registration Office , Office for National Statistics and NHS Central Register, about my NHS registration , health status , treatment and use of health services.	I agree
EDUCATION	
I authorise the English Department for Children, Schools and Families, the Welsh Department for Children, Education, Lifelong Learning, and Skills, the Scottish Government Education Directorate, or the Department of Education / Education and Skills Authority in Northern Ireland to provide information from my child's educational records to Life Study.	I agree

<<Insert Parent/Guardian Name>>	*<<Insert Date>>*	*<<Insert Parent/Guardian Signature>>*
Parent/Guardian name	**Date**	**Signature**

<<Insert Staff Member Name>>	*<<Insert Date>>*
Staff member name	**Date**

Consent form for child at 4 month visit



LIFE STUDY Understanding lives – now and for the future

Consent form for child at 4 month visit (MC)
Version 1.0
(dated 8 August 2012)

Centre Number:
Participant Identification (ID) Number:

Consent Form

Title of Project: Life Study

Thank you for taking part in Life Study. Now your child has been born, this consent form confirms that you agree to your child continuing to take part in Life Study. Please answer each of the following questions on the touch-screen and then sign the computer pad.

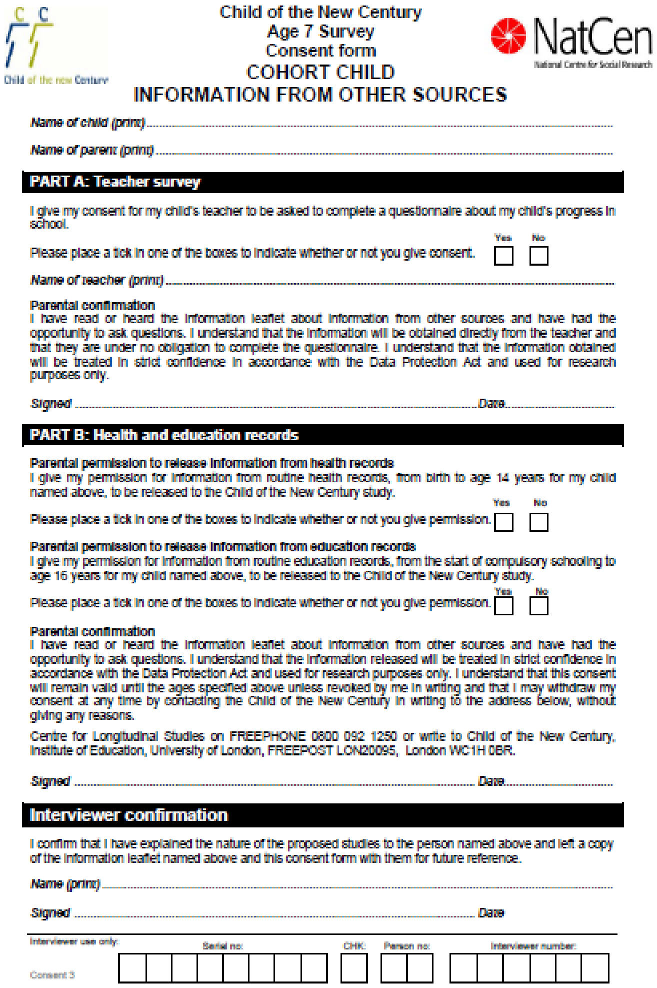
Child's name	*<<Insert Child's Name>>*

I have read and understand the Participant Information Sheet *dated 8th August 2012 (version 1.0)* for the above study. I have had the opportunity to consider the information and ask questions.	I agree
I confirm that I am the parent or legal guardian of this child.	I agree
I understand that agreeing to my child's participation in Life Study is voluntary and that I am free to withdraw consent at any time without giving any reason.	I agree
I give permission for access to my child's medical and other health-related records, and for long-term storage and use of this and other information about him/her, for health-related research purposes (even after my incapacity or death).	I agree
I give permission for individuals from regulatory authorities and the sponsoring organisation to have access to sections of my child's medical notes and data collected during Life Study, where it is relevant to taking part in this research and for monitoring and audit purposes.	I agree
I give permission for collection, long-term storage and use of my child's biological samples for health-related research purposes (even after my incapacity or death), and relinquish all rights to these samples which I am donating to Life Study.	I agree

Supplementary Figure 6 Millennium Cohort Study consent form

Millennium Cohort Study

(http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=851&sitesectiontitle=Welcome+to+the+Millennium+Cohort+Study)



Child of the New Century
Age 7 Survey
Consent form
COHORT CHILD
INFORMATION FROM OTHER SOURCES

Child of the new Century
NatCen
National Centre for Social Research

Name of child (print) ....

Name of parent (print) ....

PART A: Teacher survey

I give my consent for my child's teacher to be asked to complete a questionnaire about my child's progress in school.

Please place a tick in one of the boxes to indicate whether or not you give consent. Yes □ No □

Name of teacher (print) ....

Parental confirmation
I have read or heard the information leaflet about information from other sources and have had the opportunity to ask questions. I understand that the information will be obtained directly from the teacher and that they are under no obligation to complete the questionnaire. I understand that the information obtained will be treated in strict confidence in accordance with the Data Protection Act and used for research purposes only.

Signed ....*Date*.............................

PART B: Health and education records

Parental permission to release information from health records
I give my permission for information from routine health records, from birth to age 14 years for my child named above, to be released to the Child of the New Century study.

Please place a tick in one of the boxes to indicate whether or not you give permission. Yes □ No □

Parental permission to release information from education records
I give my permission for information from routine education records, from the start of compulsory schooling to age 16 years for my child named above, to be released to the Child of the New Century study.

Please place a tick in one of the boxes to indicate whether or not you give permission. Yes □ No □

Parental confirmation
I have read or heard the information leaflet about information from other sources and have had the opportunity to ask questions. I understand that the information released will be treated in strict confidence in accordance with the Data Protection Act and used for research purposes only. I understand that this consent will remain valid until the ages specified above unless revoked by me in writing and that I may withdraw my consent at any time by contacting the Child of the New Century in writing to the address below, without giving any reasons.

Centre for Longitudinal Studies on FREEPHONE 0800 092 1250 or write to Child of the New Century, Institute of Education, University of London, FREEPOST LON20095, London WC1H 0BR.

Signed ....*Date*.............................

Interviewer confirmation

I confirm that I have explained the nature of the proposed studies to the person named above and left a copy of the information leaflet named above and this consent form with them for future reference.

Name (print) ....

Signed .... *Date*

Interviewer use only:

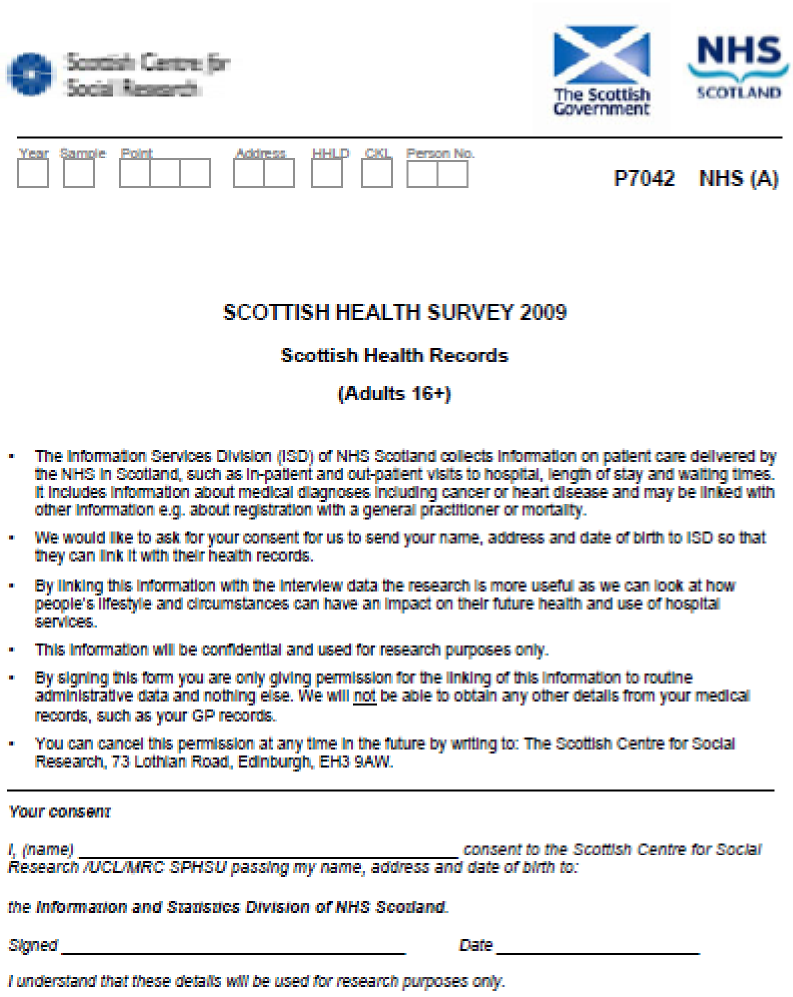
Consent 3

Serial no:	CHK:	Person no:	Interviewer number:
□□□□□□□□	□	□□	□□□□□□

Supplementary Figure 7 Scottish Health Surveys consent forms

Scottish Health Survey (http://www.gov.scot/Topics/Statistics/Browse/Health/scottish-health-survey)

Scottish Health Records – Adults 16+



Scottish Centre for Social Research

The Scottish Government

NHS SCOTLAND

Year □ Sample □ Point □□□ Address □□ HHLD □ CKL □ Person No. □□

P7042 NHS (A)

SCOTTISH HEALTH SURVEY 2009

Scottish Health Records

(Adults 16+)

- The Information Services Division (ISD) of NHS Scotland collects information on patient care delivered by the NHS in Scotland, such as in-patient and out-patient visits to hospital, length of stay and waiting times. It includes information about medical diagnoses including cancer or heart disease and may be linked with other information e.g. about registration with a general practitioner or mortality.
- We would like to ask for your consent for us to send your name, address and date of birth to ISD so that they can link it with their health records.
- By linking this information with the interview data the research is more useful as we can look at how people's lifestyle and circumstances can have an impact on their future health and use of hospital services.
- This information will be confidential and used for research purposes only.
- By signing this form you are only giving permission for the linking of this information to routine administrative data and nothing else. We will not be able to obtain any other details from your medical records, such as your GP records.
- You can cancel this permission at any time in the future by writing to: The Scottish Centre for Social Research, 73 Lothian Road, Edinburgh, EH3 9AW.

Your consent

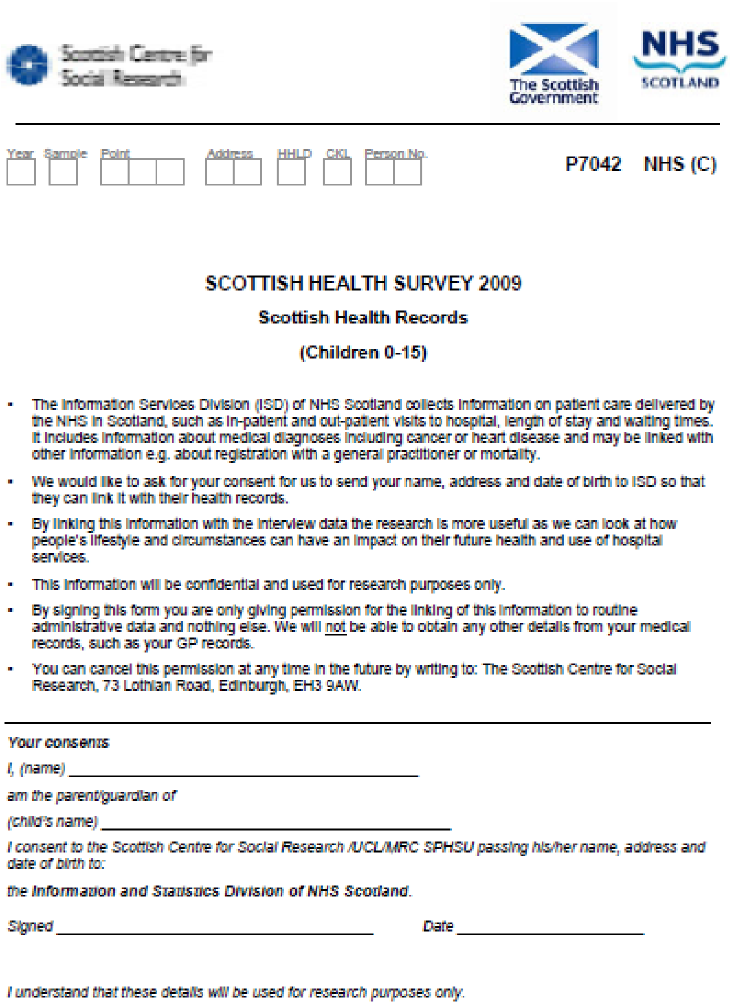
I, (name) ______________________________ consent to the Scottish Centre for Social Research /UCL/MRC SPHSU passing my name, address and date of birth to:

*the **Information and Statistics Division of NHS Scotland**.*

Signed ____________________________ *Date* _________________

I understand that these details will be used for research purposes only.

Scottish Health Records – Children 0-15



Scottish Centre for Social Research

The Scottish Government

NHS SCOTLAND

Year □ Sample □ Point □□□ Address □□ HHLD □ CKL □ Person No. □□

P7042 NHS (C)

SCOTTISH HEALTH SURVEY 2009

Scottish Health Records

(Children 0-15)

- The Information Services Division (ISD) of NHS Scotland collects information on patient care delivered by the NHS in Scotland, such as in-patient and out-patient visits to hospital, length of stay and waiting times. It includes information about medical diagnoses including cancer or heart disease and may be linked with other information e.g. about registration with a general practitioner or mortality.
- We would like to ask for your consent for us to send your name, address and date of birth to ISD so that they can link it with their health records.
- By linking this information with the interview data the research is more useful as we can look at how people's lifestyle and circumstances can have an impact on their future health and use of hospital services.
- This information will be confidential and used for research purposes only.
- By signing this form you are only giving permission for the linking of this information to routine administrative data and nothing else. We will not be able to obtain any other details from your medical records, such as your GP records.
- You can cancel this permission at any time in the future by writing to: The Scottish Centre for Social Research, 73 Lothian Road, Edinburgh, EH3 9AW.

Your consents

I, (name) ____________________________

am the parent/guardian of

(child's name) ____________________________

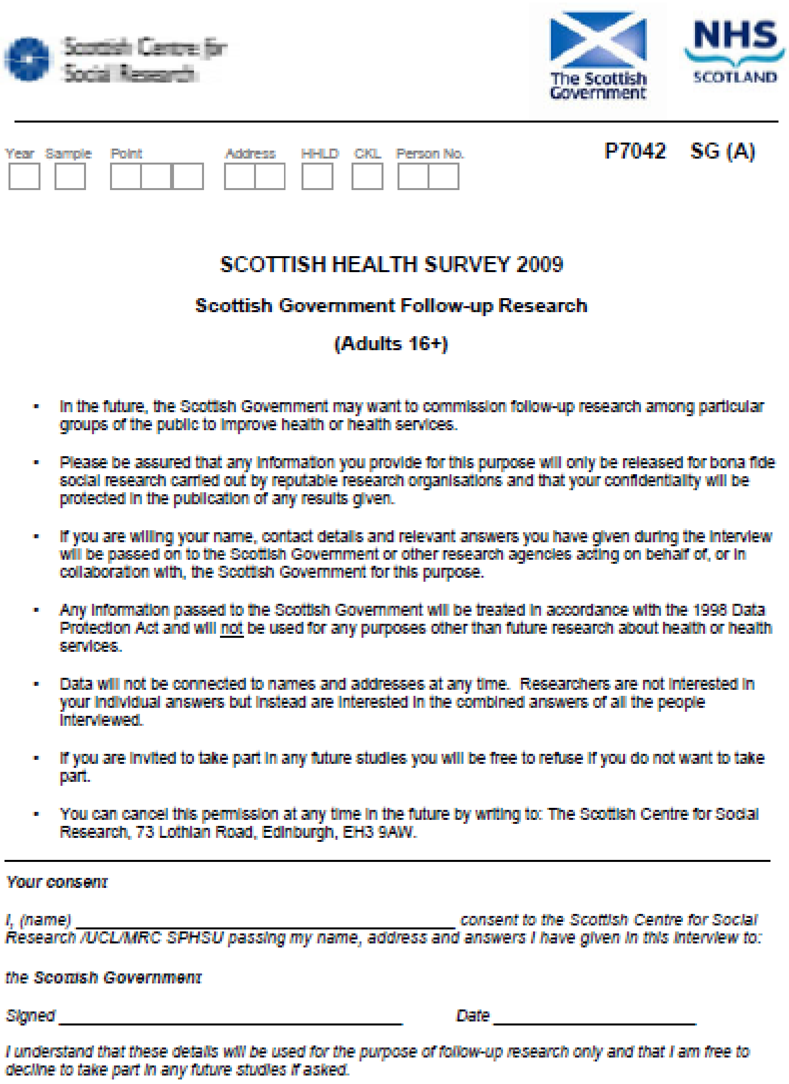
I consent to the Scottish Centre for Social Research /UCL/MRC SPHSU passing his/her name, address and date of birth to:

*the **Information and Statistics Division of NHS Scotland**.*

Signed ____________________________ *Date* _________________

I understand that these details will be used for research purposes only.

Scottish Government Follow-up Research – Adults 16+ Scottish Government Follow-up – Research children 0-15



Scottish Centre for Social Research

The Scottish Government

NHS SCOTLAND

Year	Sample	Point	Address	HHLD	CKL	Person No.
□	□	□□□	□□	□	□	□□

P7042 SG (A)

SCOTTISH HEALTH SURVEY 2009

Scottish Government Follow-up Research

(Adults 16+)

- In the future, the Scottish Government may want to commission follow-up research among particular groups of the public to improve health or health services.
- Please be assured that any information you provide for this purpose will only be released for bona fide social research carried out by reputable research organisations and that your confidentiality will be protected in the publication of any results given.
- If you are willing your name, contact details and relevant answers you have given during the interview will be passed on to the Scottish Government or other research agencies acting on behalf of, or in collaboration with, the Scottish Government for this purpose.
- Any information passed to the Scottish Government will be treated in accordance with the 1998 Data Protection Act and will not be used for any purposes other than future research about health or health services.
- Data will not be connected to names and addresses at any time. Researchers are not interested in your individual answers but instead are interested in the combined answers of all the people interviewed.
- If you are invited to take part in any future studies you will be free to refuse if you do not want to take part.
- You can cancel this permission at any time in the future by writing to: The Scottish Centre for Social Research, 73 Lothian Road, Edinburgh, EH3 9AW.

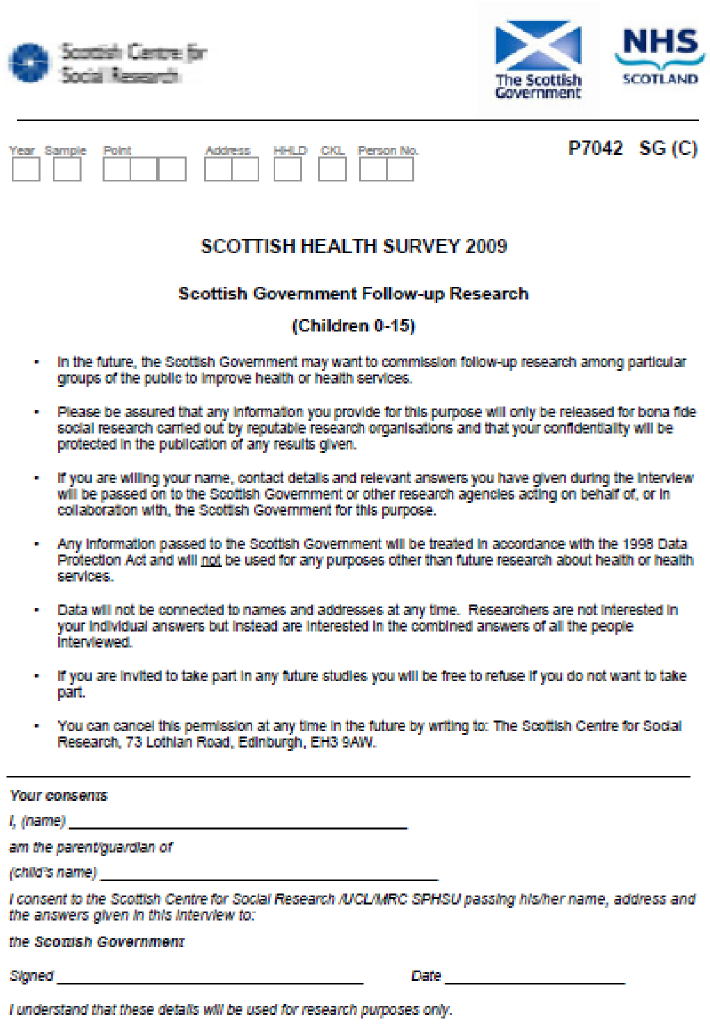
Your consent

I, (name) ____ consent to the Scottish Centre for Social Research /UCL/MRC SPHSU passing my name, address and answers I have given in this interview to:

the Scottish Government

Signed ____________________________________ *Date* _____________________

I understand that these details will be used for the purpose of follow-up research only and that I am free to decline to take part in any future studies if asked.



Scottish Centre for Social Research

The Scottish Government

NHS SCOTLAND

Year	Sample	Point	Address	HHLD	CKL	Person No.
□	□	□□□	□□	□	□	□□

P7042 SG (C)

SCOTTISH HEALTH SURVEY 2009

Scottish Government Follow-up Research

(Children 0-15)

- In the future, the Scottish Government may want to commission follow-up research among particular groups of the public to improve health or health services.
- Please be assured that any information you provide for this purpose will only be released for bona fide social research carried out by reputable research organisations and that your confidentiality will be protected in the publication of any results given.
- If you are willing your name, contact details and relevant answers you have given during the interview will be passed on to the Scottish Government or other research agencies acting on behalf of, or in collaboration with, the Scottish Government for this purpose.
- Any information passed to the Scottish Government will be treated in accordance with the 1998 Data Protection Act and will not be used for any purposes other than future research about health or health services.
- Data will not be connected to names and addresses at any time. Researchers are not interested in your individual answers but instead are interested in the combined answers of all the people interviewed.
- If you are invited to take part in any future studies you will be free to refuse if you do not want to take part.
- You can cancel this permission at any time in the future by writing to: The Scottish Centre for Social Research, 73 Lothian Road, Edinburgh, EH3 9AW.

Your consents

I, (name) ___________________________________

am the parent/guardian of

(child's name) ___________________________________

I consent to the Scottish Centre for Social Research /UCL/MRC SPHSU passing his/her name, address and the answers given in this interview to:

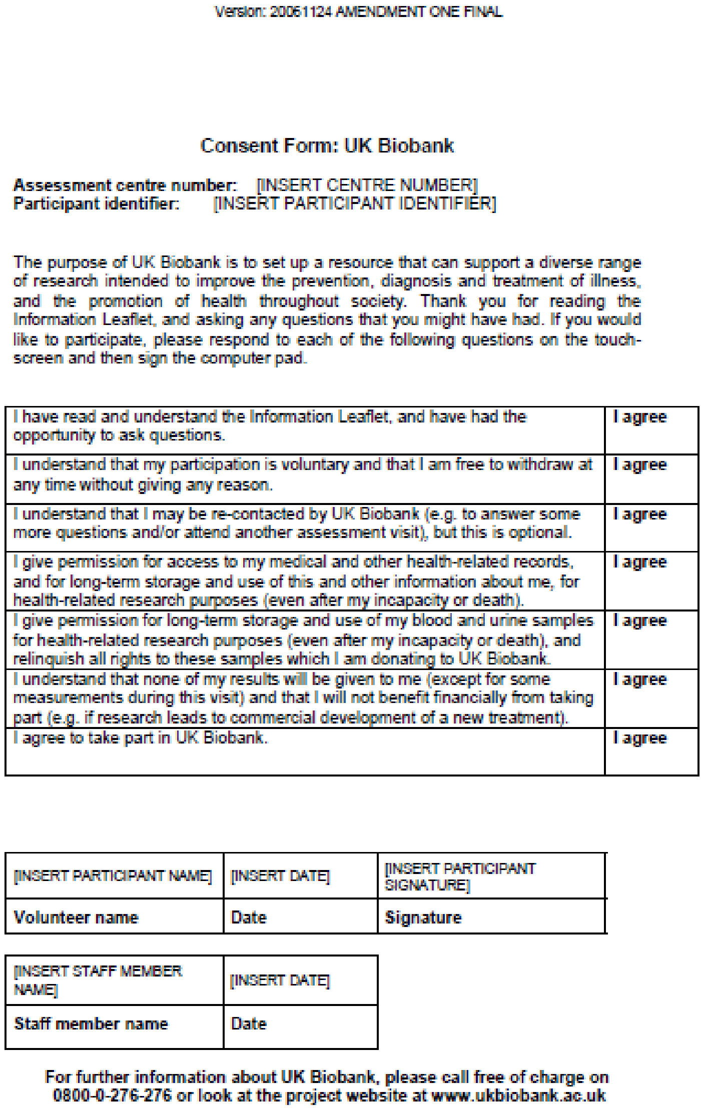
the Scottish Government

Signed ________________________________ *Date* ____________________

I understand that these details will be used for research purposes only.

Supplementary Figure 8 UK Biobank consent form

UK Biobank (https://www.ukbiobank.ac.uk/)



Version: 20061124 AMENDMENT ONE FINAL

Consent Form: UK Biobank

Assessment centre number: [INSERT CENTRE NUMBER]
Participant identifier: [INSERT PARTICIPANT IDENTIFIER]

The purpose of UK Biobank is to set up a resource that can support a diverse range of research intended to improve the prevention, diagnosis and treatment of illness, and the promotion of health throughout society. Thank you for reading the Information Leaflet, and asking any questions that you might have had. If you would like to participate, please respond to each of the following questions on the touch-screen and then sign the computer pad.

I have read and understand the Information Leaflet, and have had the opportunity to ask questions.	**I agree**
I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason.	**I agree**
I understand that I may be re-contacted by UK Biobank (e.g. to answer some more questions and/or attend another assessment visit), but this is optional.	**I agree**
I give permission for access to my medical and other health-related records, and for long-term storage and use of this and other information about me, for health-related research purposes (even after my incapacity or death).	**I agree**
I give permission for long-term storage and use of my blood and urine samples for health-related research purposes (even after my incapacity or death), and relinquish all rights to these samples which I am donating to UK Biobank.	**I agree**
I understand that none of my results will be given to me (except for some measurements during this visit) and that I will not benefit financially from taking part (e.g. if research leads to commercial development of a new treatment).	**I agree**
I agree to take part in UK Biobank.	**I agree**

[INSERT PARTICIPANT NAME]	[INSERT DATE]	[INSERT PARTICIPANT SIGNATURE]
Volunteer name	**Date**	**Signature**

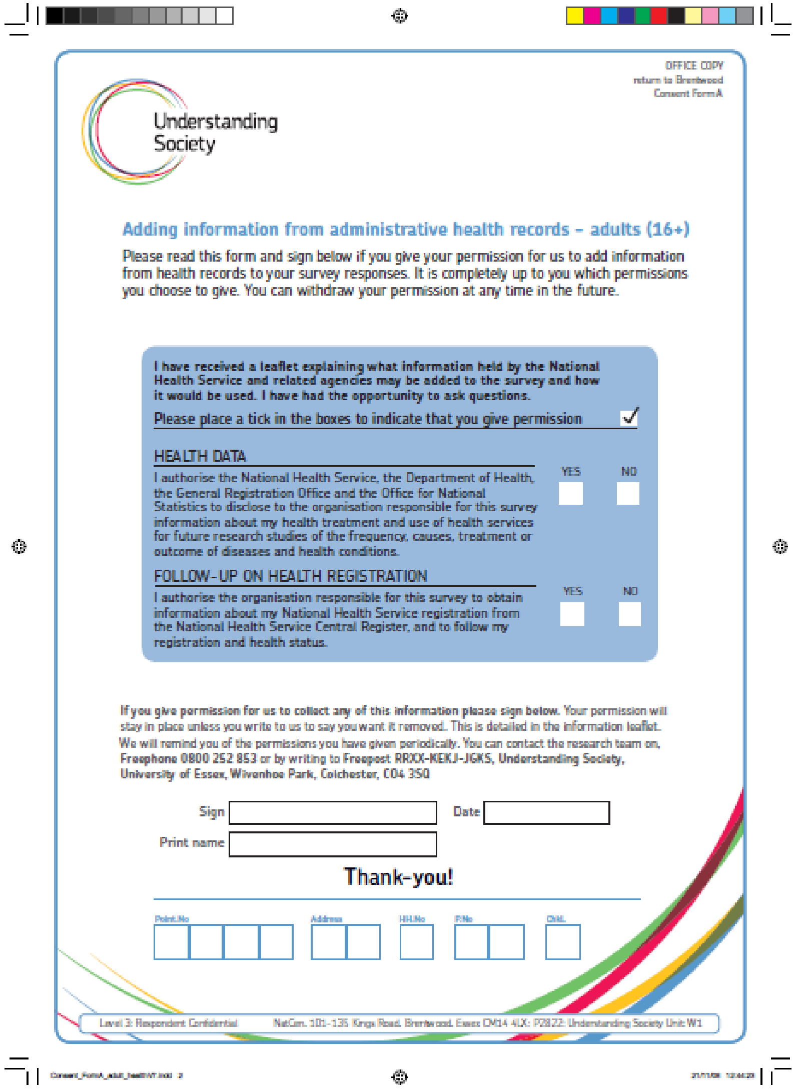
[INSERT STAFF MEMBER NAME]	[INSERT DATE]
Staff member name	**Date**

For further information about UK Biobank, please call free of charge on 0800-0-276-276 or look at the project website at www.ukbiobank.ac.uk

Supplementary Figure 9 Understanding Society consent form

Understanding Society (https://www.understandingsociety.ac.uk/)

Form A



OFFICE COPY
return to Brentwood
Consent Form A

Understanding Society

Adding information from administrative health records – adults (16+)

Please read this form and sign below if you give your permission for us to add information from health records to your survey responses. It is completely up to you which permissions you choose to give. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what information held by the National Health Service and related agencies may be added to the survey and how it would be used. I have had the opportunity to ask questions.

Please place a tick in the boxes to indicate that you give permission ✓

HEALTH DATA

I authorise the National Health Service, the Department of Health, the General Registration Office and the Office for National Statistics to disclose to the organisation responsible for this survey information about my health treatment and use of health services for future research studies of the frequency, causes, treatment or outcome of diseases and health conditions. YES □ NO □

FOLLOW-UP ON HEALTH REGISTRATION

I authorise the organisation responsible for this survey to obtain information about my National Health Service registration from the National Health Service Central Register, and to follow my registration and health status. YES □ NO □

If you give permission for us to collect any of this information please sign below. Your permission will stay in place unless you write to us to say you want it removed. This is detailed in the information leaflet. We will remind you of the permissions you have given periodically. You can contact the research team on, **Freephone 0800 252 853** or by writing to **Freepost RRXX-KEKJ-JGKS, Understanding Society, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ**

Sign □ Date □

Print name □

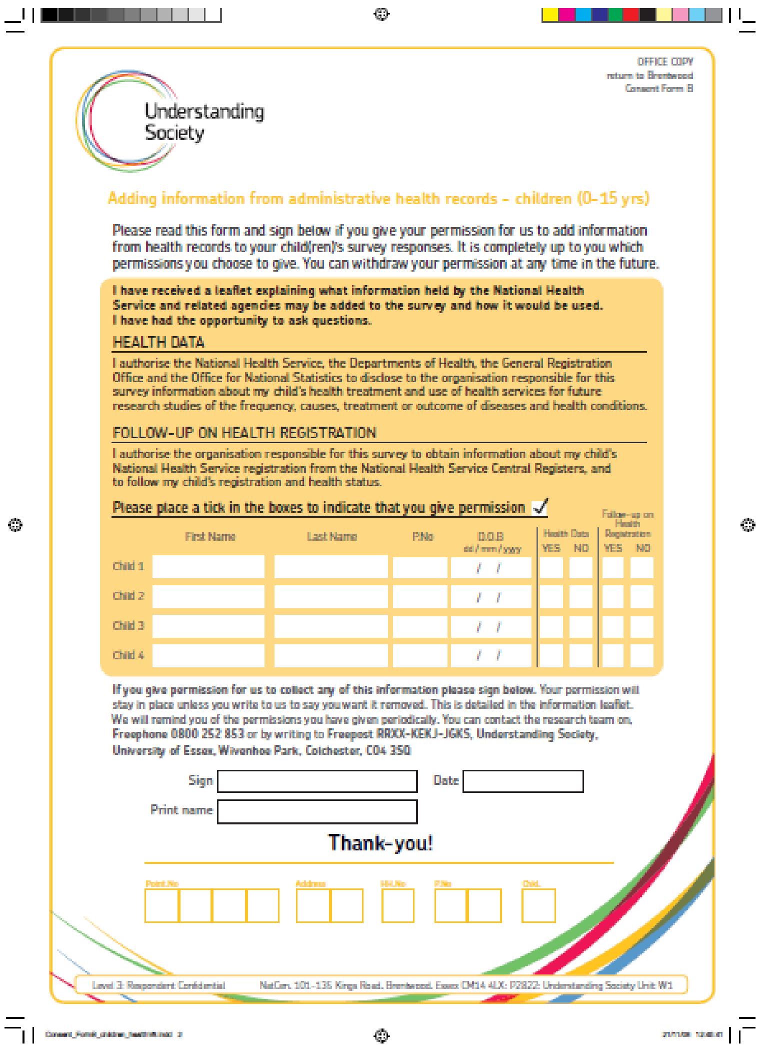
Thank-you!

Point.No □□□□ Address □□ HH.No □ P.No □□ Chkl. □

Level 3: Respondent Confidential NatCen. 101-135 Kings Road, Brentwood, Essex CM14 4LX: P2822: Understanding Society Unit W1

Consent_FormA_adult_healthW1.indd 2 21/11/08 12:44:29

Form B



OFFICE COPY
return to Brentwood
Consent Form B

Understanding Society

Adding information from administrative health records – children (0-15 yrs)

Please read this form and sign below if you give your permission for us to add information from health records to your child(ren)'s survey responses. It is completely up to you which permissions you choose to give. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what information held by the National Health Service and related agencies may be added to the survey and how it would be used. I have had the opportunity to ask questions.

HEALTH DATA

I authorise the National Health Service, the Departments of Health, the General Registration Office and the Office for National Statistics to disclose to the organisation responsible for this survey information about my child's health treatment and use of health services for future research studies of the frequency, causes, treatment or outcome of diseases and health conditions.

FOLLOW-UP ON HEALTH REGISTRATION

I authorise the organisation responsible for this survey to obtain information about my child's National Health Service registration from the National Health Service Central Registers, and to follow my child's registration and health status.

Please place a tick in the boxes to indicate that you give permission ✓

	First Name	Last Name	P.No	D.O.B dd / mm / yyyy	Health Data		Follow-up on Health Registration	
					YES	NO	YES	NO
Child 1				/ /	□	□	□	□
Child 2				/ /	□	□	□	□
Child 3				/ /	□	□	□	□
Child 4				/ /	□	□	□	□

If you give permission for us to collect any of this information please sign below. Your permission will stay in place unless you write to us to say you want it removed. This is detailed in the information leaflet. We will remind you of the permissions you have given periodically. You can contact the research team on, **Freephone 0800 252 853** or by writing to **Freepost RRXX-KEKJ-JGKS, Understanding Society, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ**

Sign □ Date □

Print name □

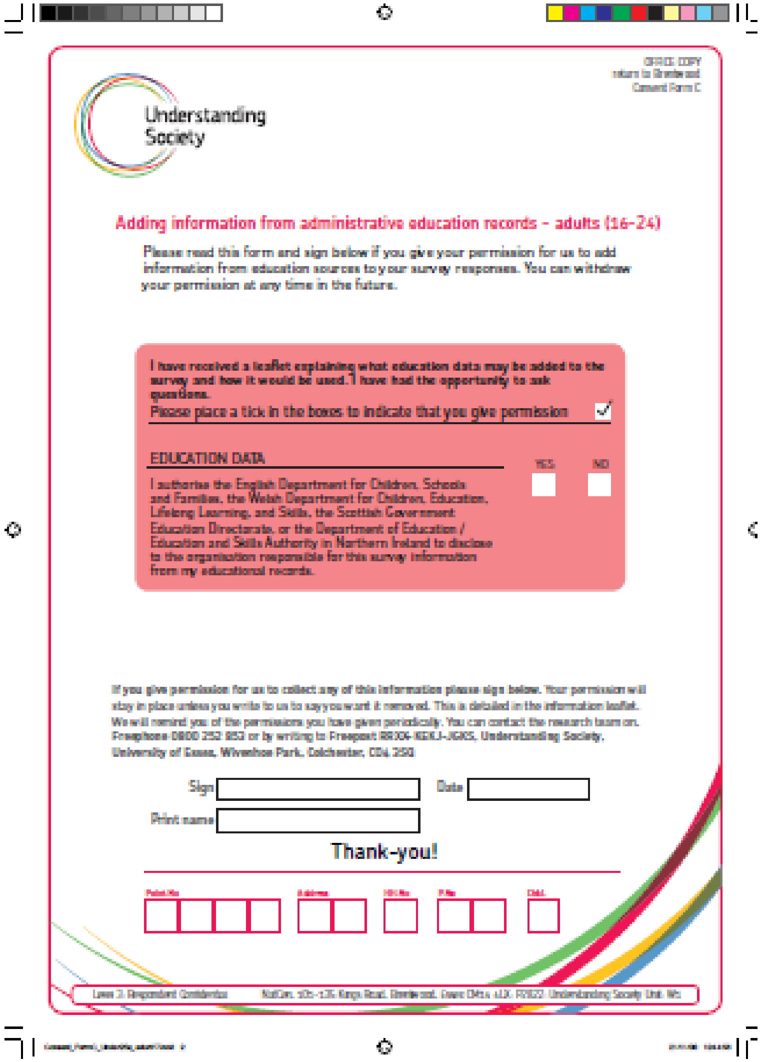
Thank-you!

Point.No □□□□ Address □□ HH.No □ P.No □□ Chkl. □

Level 3: Respondent Confidential NatCen. 101-135 Kings Road, Brentwood, Essex CM14 4LX: P2822: Understanding Society Unit W1

Consent_FormB_children_healthW1.indd 2 21/11/08 12:45:41

Form C



OFFICE COPY
return to Brentwood
Consent Form C

Understanding Society

Adding information from administrative education records – adults (16–24)

Please read this form and sign below if you give your permission for us to add information from education sources to your survey responses. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what education data may be added to the survey and how it would be used. I have had the opportunity to ask questions.

Please place a tick in the boxes to indicate that you give permission ✓

EDUCATION DATA	YES	NO
I authorise the English Department for Children, Schools and Families, the Welsh Department for Children, Education, Lifelong Learning, and Skills, the Scottish Government Education Directorate, or the Department of Education / Education and Skills Authority in Northern Ireland to disclose to the organisation responsible for this survey information from my educational records.	□	□

If you give permission for us to collect any of this information please sign below. Your permission will stay in place unless you write to us to say you want it removed. This is detailed in the information leaflet. We will remind you of the permissions you have given periodically. You can contact the research team on, **Freephone 0800 252 853** or by writing to **Freepost RRXX-KEKJ-JGKS, Understanding Society, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ**

Sign □ Date □

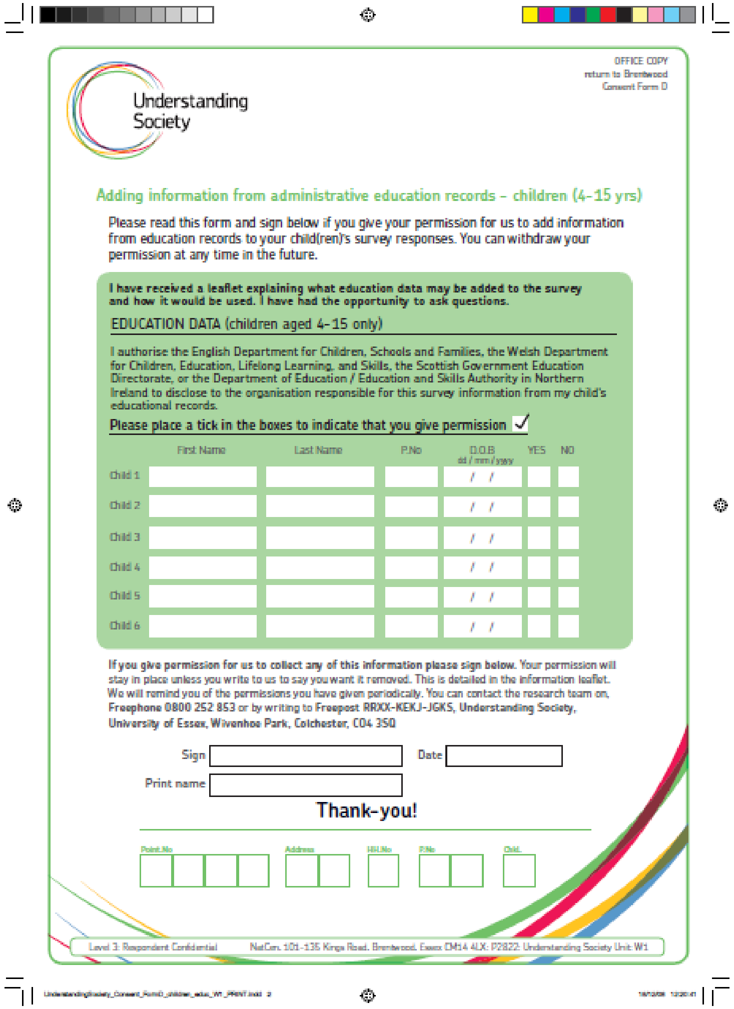
Print name □

Thank-you!

Point.No □□□□ Address □□ HH.No □ P.No □□ Chld. □

Level 3: Respondent Confidential NatCen, 101-135 Kings Road, Brentwood, Essex CM14 4LX: P2822: Understanding Society Unit W1

Dorm D



OFFICE COPY
return to Brentwood
Consent Form D

Understanding Society

Adding information from administrative education records – children (4–15 yrs)

Please read this form and sign below if you give your permission for us to add information from education records to your child(ren)'s survey responses. You can withdraw your permission at any time in the future.

I have received a leaflet explaining what education data may be added to the survey and how it would be used. I have had the opportunity to ask questions.

EDUCATION DATA (children aged 4–15 only)

I authorise the English Department for Children, Schools and Families, the Welsh Department for Children, Education, Lifelong Learning, and Skills, the Scottish Government Education Directorate, or the Department of Education / Education and Skills Authority in Northern Ireland to disclose to the organisation responsible for this survey information from my child's educational records.

Please place a tick in the boxes to indicate that you give permission ✓

	First Name	Last Name	P.No	D.O.B dd / mm / yyyy	YES	NO
Child 1				/ /		
Child 2				/ /		
Child 3				/ /		
Child 4				/ /		
Child 5				/ /		
Child 6				/ /		

If you give permission for us to collect any of this information please sign below. Your permission will stay in place unless you write to us to say you want it removed. This is detailed in the information leaflet. We will remind you of the permissions you have given periodically. You can contact the research team on, **Freephone 0800 252 853** or by writing to **Freepost RRXX-KEKJ-JGKS, Understanding Society, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ**

Sign □ Date □

Print name □

Thank-you!

Point.No □□□□ Address □□ HH.No □ P.No □□ Chld. □

Level 3: Respondent Confidential NatCen, 101-135 Kings Road, Brentwood, Essex CM14 4LX: P2822: Understanding Society Unit W1

UnderstandingSociety_Consent_FormD_children_educ_W1_PRINT.indd 2 18/12/08 12:20:41

Supplementary Table 15: ALSPAC

Records	People	Consent form	Information document
• Health records ○ Involvement in study (advising) ▪ Healthcare provider/professional ▪ GP ○ Provision of health records ○ Provision of information ▪ Locations • Databases • Records ▪ Organisations • The NHS Information Centre • NHS Central Register • Education records ○ School records ▪ Organisation • Department of Education ○ Further education ▪ Organisation • The Data Service • Department for Business, Innovation and Skills ○ Higher education ▪ Organisation • Universities and Colleges Admission Service (UCAS) • Higher Education Statistics Agency • Salary and benefits	• Person giving consent (cannot be on behalf of another)	• Organisation ○ University • Descriptions and aims • Confirmatory statements • Signature, date and printed full name	• Definitions • Benefits • How data linkage actually works • Says how consent lasts until it is withdrawn • Case studies – where record linkage has been used in the past

○ Organisations ▪ Department for Work and Pensions ▪ HM Revenue and Customs ○ Information ▪ National Insurance Contributions ▪ Tax records ▪ Pensions ▪ Savings ▪ Benefits ▪ Work and employment • Police/legal records ○ Organisation ▪ Ministry of Justice ○ Information ▪ Convictions ▪ Official cautions			

Supplementary Table 16 British Household Panel Survey

Records	People	Consent form	Information document
• Administrative health records (B) ○ Health data ○ Follow-up on health registration • From other sources ○ All adults 16-24 (C) ▪ Education data ▪ National insurance contributions, benefits and tax records, saving and pensions ○ All adults (D) ▪ National insurance contributions, benefits and tax records, savings and pensions ○ All households with a child aged 3 to 15 years (E) ▪ Education data (4-15 only)	• Administrative health records (B) ○ Parent/guardian for child • From other sources ○ All adults 16-24 (C) ▪ Person consenting for themselves ○ All adults (D) ▪ Parent/guardi an for child ○ All households with a child aged 3 to 15 years (E)	• Administrative health records (B) ○ Confirmatory permission • From other sources ○ All adults 16-24 (C) ▪ Signature, date and printed full name ▪ Confirmatory permission ○ All adults (D) ▪ Signature, date and printed full name ▪ Confirmatory permission ○ All households with a child aged 3 to 15 years (E) ▪ Signature, date and printed full name • Confirmatory permission	• Types of information • Who can use it • What does your consent cover • How long does it last • Possibility of including information relating to children under 16 • Data security • Withdrawal of consent • Contact details

Supplementary Table 17 Health Survey for England 2010

Records	People	Consent form	Information document
Hospital Episode Statistics • Adults 16+ ○ Authorisation of disclosure of HES data and linking of information NHS Central Register and Cancer Register • Adults 16+ ○ Authorisation to disclose personal information to National Health Service Central Register / to follow up health status	NHS Central Register and Cancer Register • Adults 16+ ○ Demographics ▪ Name ▪ Date of birth ○ Confirmatory Permission for person giving consent ○ Withdrawal of consent ○ People ▪ Interviewee and interviewer Hospital Episode Statistics • Adults 16+ ○ Demographics ▪ Name and signature ○ People ▪ Respondent and interviewer	NHS Central Register and Cancer Register • Adults 16+ ○ Unique identifiers ○ Organisations ▪ University ▪ Funding agency ▪ NHS Central Registrar ○ Description and aims ○ Statements ○ Dates Hospital Episode Statistics • Adults 16+ ○ Unique identifiers ○ Organisations ▪ University ▪ Funding agency ▪ NHS Information Centre ○ Description and aims ○ Dates	Could not find any accompanying documentation

Supplementary Table 18 Millennium Cohort Study

Records	People	Consent form	Information document
Child of the New Century Age 7 Survey • Information from other sources ○ Teacher Survey • Consent to contact teacher ○ Health and education records	• Teacher ○ Confirmatory permission • Parent/guardian ○ Parental permission to release of information from health records ○ Parental permission to release information education records ○ Confirmatory permission • Interviewer ○ Interviewer confirmation • Confirmatory permission • Child (for whom consent is being given)	• Interviewer confirmation • Organisations • Form is divided into sections specifically for the teacher, parent/guardian and interviewer • Consent statements • Description/aims • Unique identifiers	• Explains what information from other sources is • Explain might contact people from outside the family – school teachers • Explain about collecting information from routine records on education • Explain about collecting information from routine medical and other health related records • Explain about collecting information from routine records of economic circumstances

Supplementary Table 19 Scottish Health Survey

Records	People	Consent form	Information document
Scottish Health Records • Adults 16+ • Consent to send information to allow health record linkage o Name o Address o Date of birth • Children 0-15 • Consent to send information to allow health record linkage – on behalf of child o Name o Address o Date of birth Scottish Government Follow-up Research • Adults 16+ • Consent to pass information to the Scottish centre for Social Research o Name o Address o Relevant answers • Children 0-15 • Consent to pass information to the Scottish centre for Social Research – on behalf on child o Name	Scottish Health Records • Adults 16+ o Person giving consent • Children 0-15 o Parent/guardian o Child Scottish Government Follow-up Research • Adults 16+ o Person giving consent • Children 0-15 o Parent/guardian o Child	• Organisations o Government o NHS Scotland • Unique identifiers • Consent statements • Confirmation of understanding	• What is the survey • Who takes part • What are the questions about • General fact – “did you know…” • What can participants find out more • How does the Scottish government use the information • Who else uses the information • Contact details Ask all 16+ • NHSCanA o Consent for linkage with health records o Associated documentation ▪ Pale green consent form • ReInterA o Consent for use of personal information as part of further studies o Associated documentation ▪ Pale blue consent form Ask all aged 13-15 • NHSCanY o Consent for linkage with health records – given by child o Associated documentation ▪ Lemon consent form • ReInterY o Consent for use of personal information as part of further studies – given by child o Associated documentation ▪ Pink consent form Ask all aged 0-13 • NHSCanC/NHSCanY o Consent for linkage with health records –

○ Address ○ Relevant answers			given by respondent on behalf of child/children ○ Associated documentation ▪ Lemon consent form • NHSCon ○ Coding/categories ▪ 1 – Consent given ▪ 2 – Consent not given ○ IF NHSCon = Consent given THEN • NHSSig ○ Consent for use of personal information as part of further studies – given by respondent on behalf of child/children ○ Associated documentation ▪ Pink consent form • ReIntCon ○ Coding/categories ▪ 1 – Consent given ▪ 2 – Consent not given ○ IF ReIntCon = Consent given THEN • eIntSigritten consent

Supplementary Table 20 UK Biobank

Records	People	Consent form	Information document
• Records ○ Status ▪ Living ▪ Incapacitated ▪ Dead ○ Access to… ▪ Medical ▪ Other health related ○ Storage ▪ Long-term ○ Use – health-related research purposes ▪ Information • Provided information • Other information • Samples ○ Storage ▪ Long-term ○ Samples given ▪ Blood ▪ Urine ○ Rights ▪ Relinquished to UK Biobank • Results ○ Access to results ▪ None given; although certain may be shared during a visit ○ Benefits ▪ No financial benefit	• Volunteer • Staff member • Participation ○ Type ▪ Voluntary ○ Revocation (without reason)	• Description and aims • Consent statement • Date and signatures • Unique identifier • Agreements • Contact (optional) ○ Answer further questions ○ Attend another assessment visit	• Purpose of Biobank • Why were people chosen – used DOB to check age and told practices that there patients were being invited to a study • What does being in Biobank involve • What happens during a visit • What happens after a visit • What to do if you do not/want to take part • What to do if you are not sure • What to do before the assessment visit/preparation • Travel expenses • Why do you need consent • Do the participants need to agree to everything • Are there any risks • How will the information about me be kept confidential • Who will be able to use my information • Withdrawal- levels • What happens if something goes wrong – study has insurance • Funding/organising study • Contact details

Supplementary Table 21 Understanding Society

Records	People	Consent form	Information document
Consent Form A • Adult Health form ○ Adding information from administrative health records – adults 16+ ▪ Health data • Health treatment • Use of health services • Future research studies, inclusive of ○ Frequency ○ Causes ○ Treatment or outcome of diseases ○ Health conditions ▪ Follow-Up Health Registration • National Health Service Central Register ○ National Health Service registration ▪ Registration ▪ Health status Consent Form B • Child Health form ○ Adding information from administrative health records – children (0-15yrs)	Consent Form A • Adult Health form ○ Adding information from administrative health records – adults 16+ ○ Person giving consent Consent Form B • Child Health form ○ Adding information from administrative health records – children (0-15yrs) ○ Parent/guardian and child/ren Consent Form C • Under 25 Education Form ○ Adding information from Administrative education Records – adults (16-24) ▪ Person giving consent Consent Form D • Child Education form ○ Adding information from administrative education records – children (4-15yrs) ○ Parent/guardian and child/ren	Consent Form A • Adult Health form ○ Adding information from administrative health records – adults 16+ • Organisations ○ National Health Service ○ Department of Health ○ General Registration Office ○ Office for National Statistics Consent Form B • Child Health form ○ Adding information from administrative health records – children (0-15yrs) • Organisations ○ National Health Service ○ Department of Health ○ General Registration Office ○ Office for National Statistics Consent Form C • Under 25 Education Form ○ Adding information from Administrative education Records – adults (16-24) • Organisations ○ England - Department for Children, Schools	• Introduction • What information would be added • Who will use it • What does the permission cover • How long does it last • Children • Data security • Changing minds • Thank you • Contact details

▪ Health data • Health treatment • Use of health services • Future research studies, inclusive of ○ Frequency ○ Causes ○ Treatment or outcome of diseases ○ Health conditions ▪ Follow-Up Health Registration • National Health Service Central Register Consent Form C • Under 25 Education Form ○ Adding information from Administrative education Records – adults (16-24) ▪ Education data ▪ Education data (records) Consent Form D • Child Education form ○ Adding information from administrative education records – children (4-15yrs) ▪ Education data (records)		and families ○ Wales - Department for Children, Education, Lifelong Learning, and Skills ○ Scotland - Government Education Directorate ○ Northern Ireland - Department of Education/Education and Skills Authority Consent Form D • Child Education form ○ Adding information from administrative education records – children (4-15yrs) • Organisations ○ England - Department for Children, Schools and families ○ Wales - Department for Children, Education, Lifelong Learning, and Skills ○ Scotland - Government Education Directorate ○ Northern Ireland - Department of Education/Education and Skills Authority	

Supplementary Table 22 Born in Bradford

Records	People	Consent form	Information document
Mechanisms of the Development of ALLergy Mothers consent form • Blood sample ○ Take child's blood sample, store for future use, may be used outside of the UK – can withdraw consent • Skin allergy test ○ Agree for child to be tested for allergies Born in Bradford Allergy and Infection Study Mother's consent form • Information collected in this study will be linked to the Born in Bradford, can contact mother when child is 4 for an allergy test • Agree to a blood sample being taken – stored for use both in and out of UK, will not be told results about infection status Mother's consent form • Agree to future contact • Agree to given biological samples – blood, saliva, urine (same basis as above) • Agree to blood sample from umbilical cord being taken post-delivery – the baby's and mother's biological samples will be stored. Understand that researchers in and out of the UK can use these samples and that the results of which will not be provided	Mechanisms of the Development of ALLergy Mothers consent form • Parent/guardian and child • Confirm understanding that researchers based at BiB, with UK, in and out of Europe may use the samples and the results of which will not be given to the mother • Name of participant, name of person taking consent if different to the researcher, name of researcher – all have date of recruitment and signature • If withdraw, agree for the retention of the samples unless mother specifically requests they are destroyed – research team will make every effort to do so Born in Bradford Allergy and Infection Study Mother's consent form • Confirmation of understanding of information and had opportunity to ask questions, participation is voluntary, free to withdraw at any time without giving any reason, without our medical care or legal rights being affected • Wish to withdraw from the study, samples are to be destroyed and research team to kame every effort that no further analysis is conducted on the samples Mother's consent form • Confirmation of understanding, participation is voluntary, can withdraw without reason • Understand that researchers may wish to	• Confirmation of understanding • Withdrawal of consent • Agree to GP knowing about child's involvement in the study • Permission to access parent/guardian and child's medical notes and data collected during the study, regulatory authorities or form the NHS Trust	• Aim of study • The process • What will happen to the information • What will happen to the samples • Contact • Confidentiality • Advantages for taking part • Will the participants get to know the findings of the research

Father's consent form - Agree to future contact - Agree to provide biological samples which may include saliva. Can withdraw from study but samples will be retained unless request to destroy all samples is made	look at medical and educational records – understand that certain other countries do not have the same data protection laws as the UK but the mother has understood that the information cannot be linked back to her - Child details may be passed to the ONS and this will need permission from the Data Advisory Group - Name of participant, name of person taking consent if different from study administrator and name of study administrator Father's consent form - Confirmation of understanding, participation is voluntary, can withdraw without reason - understand that certain other countries do not have the same data protection laws as the UK but the father has understood that the information cannot be linked - Understand that researchers in and out of the UK can use these samples and that the results of which will not be provided - Name of participant, name of person taking consent if different from study administrator and name of study administrator		

Supplementary Table 23 Life Study

Records	People	Consent form	Information document
Consent form for record linkage at 4months (child) • Health • Education Consent form for record linkage at first visit/contact (father/partner) • Health • Education • Mobile • Economic ○ Benefit claims ○ Time on employment programs • Work and employment Consent form for record linkage at first visit/contact (mother) • Health • Education • Mobile • Economic	All Forms • Confirmation of understanding and opportunity to ask questions • Confirmation of understanding that participation is voluntary and that consent may be withdrawn at any time without reason • medical notes and data – consenting for themselves or on behalf of a child • Agree to GP being informed of participation • Agree to baby(ies) taking part • Give permission for regulatory authorities and sponsoring organisations to have access Forms for child: • Contact form for child at 4 month visit • Confirmation of parent/legal guardian of child • Give permission for collection, long-term storage of child's biological samples for health related research purposes and relinquish all right to these sample which I am donating to Life Study • Consent form for child at 4 month visit (MC) • Give permission for my child to be recorded (e.g. camera) and for these recording to be stored long-term and used for research purposes • Understand that none of the child's results will be shared apart from a	All forms • Organisations ○ Health ▪ NHS ▪ DH ▪ Registers ▪ General registration ▪ ONS ▪ NHS Central Register ○ Education ▪ English Department for Children, Schools and Families ▪ Welsh Department for Children, Education, Lifelong Learning and Skills ▪ Scottish Government Education Directorate ▪ Department of Education/Education and Skills Authority in Northern Ireland Consent form for record linkage at first visit/contact (father/partner) • Organisations ○ Department for Work and Pensions ▪ HM Revenue and Customs Consent form for record linkage	• What is life study going to do • Why have I been chosen • What will happen when I take part • What happens after the visits • What happens at the pregnancy visit • What will happen to the baby • Will they get any results • Do I have to take part • Travel expenses • Benefits to mother/child by taking part • Risks • Confidentiality • What happens if there are any problems • What to do if sure/unsure/undecided about taking part • Will GP know I'm taking part • Withdrawal • Who is organising/funding study • Who has reviewed the study • Contact details • Who will use the information • When are samples taken • Record linkage ○ What is it ○ How does it work ○ Why do you need the records

	select number of measurement Forms for partner • Consent form for partner at 4month visit • May be contacted in the future • Give permission for collection, long-term storage of biological samples for health related research purposes and relinquish all right to these sample which I am donating to Life Study • Consent form for partner at 28 week visit (MC) • May be contacted in the future • Give permission for collection, long-term storage of biological samples for health related research purposes and relinquish all right to these sample which I am donating to Life Study Forms for mother (including pregnancy) • Consent form for mother at 4month visit (NC) • Consent form for pregnant mother at 28week visit (MC) • May be contacted in the future – mother and child(ren) • Access to medical and other health-related records, long-term storage • Only certain results will be given to me	at first visit/contact (mother) • Organisations ○ Education ▪ Universities and Colleges Admissions Service/Higher Education Statistics Agency (HESA) – education records ▪ Department for Business, Innovation & Skills to provide education records ○ Economic ▪ Department for Work and Pensions ▪ HM Revenue and Customs ▪ Work and employment	○ Who do you link to my child's records ○ Who will use the information ○ Which records will be linked

Elements highlighted in red indicate that an element was not found in DDI3.2 to map with. Elements highlighted in purple indicate the ‘Note’ element in DDI can only be used in conjunction with another maintainable object; otherwise, the consent element cannot be mapped. The use of the ‘Note’ element here functions very much as a work around solution.

Supplementary Table 24 Personal records cross-walk

Personal records	
Personal records analyses	**Data Documentation Initiative 3.2 (XML source)**
Health	
□ Organisation	
o National Health Service	<xs:element name="Organization" type="OrganizationType"/>
□ The NHS Information Centre	<xs:element name="OrganizationName" type="OrganizationNameType"/>
□ NHS Central Registrar	
o Department of Health	<xs:element name="Organization" type="OrganizationType"/>
o General Registration Office	
o Office for National Statistics	
□ Healthcare professional	
o Primary care	<xs:element name="Concept" type="ConceptType"/>
□ GP	<xs:element name="Contributor" type="ContributorType"/>
o Secondary care	<xs:element name="Concept" type="ConceptType"/>
o Tertiary care	
□ Clinical	
o Terminologies	<xs:element name="CodeListGroup" type="CodeListGroupType"/>
□ ICD-10	<xs:element name="CodeList" type="CodeListType"/>
□ ICD for Oncology	

□ SNOMED-CT	
□ Read Codes	
□ DSM	
□ OPCS	
□ Treatments and management of conditions	
o Current	
□ Health treatment	
□ Use of health services	
o Previous	
□ Health treatment	
□ Use of health services	
o Samples provided	
□ Method	<xs:element name="CodeListGroup" type="CodeListGroupType"/>
□ Invasive	<xs:element name="CodeList" type="CodeListType"/>
□ Non-invasive	
□ Type	<xs:element name="CodeListGroup" type="CodeListGroupType"/>
□ Blood	<xs:element name="CodeList" type="CodeListType"/>
□ Urine	
□ Hair	
□ Saliva	
□ Storage of samples	
□ Rights	
□ Benefits and compensation	
o Tests and assessments	
□ Rights to results	

□ Length of test	
□ Location	
□ Follow-up on health registration	
Education	
□ Type	<xs:element name="Organization" type="OrganizationType"/>
o School records	<xs:element name="OrganizationName" type="OrganizationNameType"/>
o Further education	
o Higher education	
□ Provider	
o Organisation	<xs:element name="Organization" type="OrganizationType"/>
□ Department of Education	<xs:element name="OrganizationName" type="OrganizationNameType"/>
□ The Data Service	
□ Department for Business, Innovation and Skills	
□ Universities and Colleges Admission Service (UCAS)	
□ Higher Education Statistics Agency	
□ Department for Children, Schools and families	
□ Department for Children, Education, Lifelong Learning, and Skills	
□ Government Education Directorate	
□ Department of Education/Education and Skills Authority	
□□□ Location	<xs:element name="LocationName" type="LocationNameType"/>
o Educators	
□ Name	<xs:element name="Contributor" type="ContributorType"/>

□ Associated school/college/university	<xs:element name="OrganizationName" type="OrganizationNameType"/>
Criminal	
□ Organisations	<xs:element name="Organization" type="OrganizationType"/>
o Ministry of Justice	<xs:element name="OrganizationName" type="OrganizationNameType"/>
□ Records	
o Official cautions	
o Convictions	
Work and employment	
□ Organisations	<xs:element name="Organization" type="OrganizationType"/>
o Department for Work and Pensions	<xs:element name="OrganizationName" type="OrganizationNameType"/>
o HM Revenue and Customs	
□ Records	
o Salary	
o National insurance contributions	
o Tax	
o Savings	
o Benefits	
o Pensions	
Mobile	
□ Past	
□ Current	
Future	

Supplementary Table 25 People cross-walk

People	
People analyses	**Data Documentation Initiative 3.2 (XML source)**
· Identifiers	
o NHS number	<xs:element name="Note" type="NoteType"/>
o Passport number	
o other	<xs:element minOccurs="0" name="FormNumber" type="xs:string"/>
· Name of interviewee	
o Forename	<xs:element minOccurs="0" name="FullName" type="r:InternationalStringType"/>
o Surname	
· Address	<xs:element name="Note" type="NoteType"/>
· Birth details	
o Date	<xs:element name="Date" type="DateType"/>
o Location	<xs:element name="LocationName" type="LocationNameType"/>
· Ethnicity	<xs:element name="Code" type="CodeType"/>
· Nationality	<xs:element name="Country" substitutionGroup="CountryCode" type="CountryType"/>
· Disability	
· Contact details (telephone number etc.)	<xs:element name="TelephoneNumber" type="xs:string"/> <xs:element name="Email" type="r:EmailType"/>
· Next of kin	<xs:element minOccurs="0" name="FullName" type="r:InternationalStringType"/>
· Family	
o Partner	<xs:element minOccurs="0" name="FullName" type="r:InternationalStringType"/>
o Dependents	
· Participant	

o Individual consenting for themselves	<xs:element name="Note" type="NoteType"/> <xs:complexType name="NoteType"> <xs:sequence> <xs:element minOccurs="0" ref="TypeOfNote"/> <xs:element minOccurs="0" ref="NoteSubject"/> <xs:element maxOccurs="unbounded" ref="Relationship"/> <xs:element minOccurs="0" name="Responsibility" type="xs:string"/> <xs:element minOccurs="0" ref="Header"/> <xs:element minOccurs="0" ref="NoteContent"/> <xs:element minOccurs="0" ref="ProprietaryInfo"/> </xs:sequence> <xs:attribute ref="xml:lang" use="optional"/> </xs:complexType>
o Parent/guardian on behalf of a child	
· Persons present	
o Date	<xs:element name="Date" type="DateType"/>
o Location	<xs:element name="LocationName" type="LocationNameType"/>
· Interviewer	
o Name	<xs:element minOccurs="0" name="FullName" type="r:InternationalStringType"/>
Forename	
Surname	
· Witnesses	
o Name	<xs:element minOccurs="0" name="FullName" type="r:InternationalStringType"/>
Forename	
Surname	

Supplementary Table 26 Consent form cross-walk

Consent form	
Consent form analyses	**Data Documentation Initiative 3.2 (XML source)**
Description/aims	
Method of collection	<xs:element name="ModeOfCollection" type="ModeOfCollectionType"/>
o CAPI	<xs:complexType name="ModeOfCollectionType"> <xs:complexContent> <xs:extension base="r:IdentifiableType"> <xs:sequence> <xs:element minOccurs="0" ref="TypeOfModeOfCollection"/> <xs:element minOccurs="0" ref="r:Description"/> </xs:sequence> </xs:extension> </xs:complexContent> </xs:complexType>
o CASI	
Undertakings	
• Declaration	<xs:element name="Note" type="NoteType"/> <xs:complexType name="NoteType"> <xs:sequence> <xs:element minOccurs="0" ref="TypeOfNote"/> <xs:element minOccurs="0" ref="NoteSubject"/> <xs:element maxOccurs="unbounded" ref="Relationship"/> <xs:element minOccurs="0" name="Responsibility" type="xs:string"/> <xs:element minOccurs="0" ref="Header"/> <xs:element minOccurs="0" ref="NoteContent"/> <xs:element minOccurs="0" ref="ProprietaryInfo"/> </xs:sequence> <xs:attribute ref="xml:lang" use="optional"/> </xs:complexType>
o Rights of participants	

Organisations	<xs:element name="Organization" type="OrganizationType"/>
o Funding agencies	<xs:element name="FundingInformation" type="FundingInformationType"/>
o Universities	<xs:element name="Organization" type="OrganizationType"/>
o Governments	
o Archive	
Questions/consent statements	
o Logic	<xs:element name="QuestionSequence" type="QuestionSequenceType"/>
o Purpose	<xs:element name="QuestionIntent" type="r:StructuredStringType"/>
o Potential Reponses	<xs:element name="ResponseText" type="DynamicTextType"/> <xs:element name="CodeList" type="CodeListType"/> <xs:element name="CodeListName" type="r:NameType"/> <xs:element name="CodeListReference" type="ReferenceType"/>
o Codes and categories	<xs:element name="CodeList" type="CodeListType"/> <xs:element name="CodeListName" type="r:NameType"/> <xs:element name="CodeListReference" type="ReferenceType"/>
Confirmatory information	
o Confirmation of understanding	<xs:element name="Note" type="NoteType"/> <xs:complexType name="NoteType"> <xs:sequence> <xs:element minOccurs="0" ref="TypeOfNote"/> <xs:element minOccurs="0" ref="NoteSubject"/> <xs:element maxOccurs="unbounded" ref="Relationship"/> <xs:element minOccurs="0" name="Responsibility" type="xs:string"/> <xs:element minOccurs="0" ref="Header"/> <xs:element minOccurs="0" ref="NoteContent"/> <xs:element minOccurs="0" ref="ProprietaryInfo"/> </xs:sequence> <xs:attribute ref="xml:lang" use="optional"/> </xs:complexType>
o Signature	
o Date	<xs:element name="Date" type="DateType"/>

o Full name	<xs:element minOccurs="0" name="FullName" type="r:InternationalStringType"/>

Supplementary Table 27 Information document cross-walk

Information document	
Information document analyses	**Data Documentation Initiative 3.2 (XML source)**
· Study	
o Aims	<xs:element name="Citation" type="CitationType"/>
o Objectives	
o Funding bodies	<xs:element name="FundingInformation" type="FundingInformationType"/>
o Reviewers	<xs:element name="Contributor" type="ContributorType"/>
o Contact details	<xs:element name="TelephoneNumber" type="xs:string"/> <xs:element name="Email" type="r:EmailType"/>
· Participation	
o Invitation process	<xs:element name="EventType" type="CodeValueType"/> <xs:element name="LifecycleEvent" type="LifecycleEventType"/>
o Benefits and risks	
Immediate	
Future	
o Consent process	
Coverage	<xs:element name="Coverage" type="CoverageType"/>
Length of time	<xs:element name="temporal" substitutionGroup="dc:coverage"/>
Withdrawal	<xs:element name="EventType" type="CodeValueType"/> <xs:element name="LifecycleEvent" type="LifecycleEventType"/>
· Levels of withdrawal	
o Visits	<xs:element name="CodeListGroup" type="CodeListGroupType"/>
Prior	<xs:element name="CodeList" type="CodeListType"/>

· Preparation	
During	<xs:element name="CodeList" type="CodeListType"/>
· Biological samples	
o Specify which ones	
o How will these be taken	
o Who will the samples be taken from	
· Questionnaires to complete	
Post	<xs:element name="CodeList" type="CodeListType"/>
· Obtaining certain results	
o Other people	
GP	<xs:element name="Contributor" type="ContributorType"/>
School teachers	
o Expenses	
travel	
· Record linkage	
o Definition	
o How is it achieved	
o Case studies/examples	
o Data	
Which records/registries will be linked to	
How will the data be accessed	
· Subsequent research	
o Who will have access to the data	

o Getting to know results	
· Confidentiality and security	
o What safeguards are in place to protect participant confidentiality and data security	

Supplementary Code 1 XML schema of combined consent elements

```
<?xml version="1.0" encoding="utf-8"?>
    <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
    <xs:element name="consent form">
        <xs:complexType>
            <xs:sequence>
                <xs:element    name="academicInstitutions"    type="xs:string"    minOccurs="1"
maxOccurs="1"/>
                <xs:element name="aim" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="archive" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="consentModel" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="fundingBodies" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element    name="governanceFrameworks"    type="xs:string"    minOccurs="1"
maxOccurs="1"/>                                        <xs:element    name="methodOfCollection"
type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="questions" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="regulatory" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="temporal" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="undertakings" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="version" type="xs:string" minOccurs="1" maxOccurs="1"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>

    <xs:element name="participant information document">
        <xs:complexType>
            <xs:sequence>
```

```
                    <xs:element name="study" type="xs:string" minOccurs="1" maxOccurs="1"/>
                    <xs:element name="confidentialityAndSecurity" type="xs:string" minOccurs="1"
        maxOccurs="1"/>
                    <xs:element name="general" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="participationProcess" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="recordLinkage" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="visits" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="questionnaires" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="consentProcess" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="research" type="xs:string" minOccurs="1" maxOccurs="1"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>

    <xs:element name="person">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="confirmationOfUnderstanding" type="xs:string" minOccurs="1"
maxOccurs="1"/>
                <xs:element name="contactDetails" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="forename" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="surname" type="xs:string" minOccurs="1" maxOccurs="1"/>
                <xs:element name="uniqueIdentifier" type="xs:string" minOccurs="1" maxOccurs="1"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>

    <xs:element name="records">
```

```
            <xs:complexType>
                <xs:sequence>
                    <xs:element name="health" type="xs:string" minOccurs="1" maxOccurs="1"/>
                    <xs:element name="education" type="xs:string" minOccurs="1" maxOccurs="1"/>
                    <xs:element name="legal" type="xs:string" minOccurs="1" maxOccurs="1"/>
                    <xs:element name="family" type="xs:string" minOccurs="1" maxOccurs="1"/>
                    <xs:element name="mobile phone usage" type="xs:string" minOccurs="1" maxOccurs="1"/>
                    <xs:element name="economic" type="xs:string" minOccurs="1" maxOccurs="1"/>
                </xs:sequence>
            </xs:complexType>
        </xs:element>
    </xs:schema>
```

Supplementary Code 2 HTML report

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
	<head>
		<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
		<title>Consent form v1.0</title>
		<link href="css/ea.css" rel="stylesheet" type="text/css" />
		<script language="JavaScript" src="js/displayToc.js" type="text/javascript"></script>
	</head>
<body onload="initLoad(this,'toc.htm','./EARoot/EA1.htm')" onresize="resizePage()">
	<div class="IndexTitle">
		Consent form v1.0
	</div>
	<div class="IndexHeader" id="IndexHeader">
		<img src="images/ea.gif" align="right" alt="Enterprise Architect" />
	</div>
	<noscript>
		<div class="NoScript">
			It appears that you may have Javascript disabled.
		</div>
	</noscript>
</body>
</html>
```

Appendix D: Grant proposal - Development a global metadata registry for epidemiological and public health research datasets derived from observational studies

Project description: The aim of this project is to develop a novel registry to collate metadata for observational study research datasets. Currently, there is currently no known mandatory registration process for this kind of study so the registry would be the first of its kind. In essence, researchers and other stakeholders involved in observational studies would need to register their data by creating a metadata record. The intention is to build a public facing, searchable platform from which members of the academic community, public and study participants may in the short term discover and learn more about certain datasets. The longer term goal is to encourage stakeholders to collaborate even more as part of an enhanced research environment aiming to increase the repurposing and reuse of research data consistent with the philosophy of the RDL.

The project will build on findings from the enhancing data discoverability study (Chapter 3) and directly address the recommendation develop a public health portal for the registering of observational studies worldwide (chapter **Error! Reference source not found.**). It will also partly address recommendation statements, 2: improve awareness of the implications associated with poor quality metadata; 4: investigate mechanisms to further integrate SWTs; and 6: increase identification of commonalities and links between studies through improved provision of openly available metadata and other associated research artefacts, all within epidemiological and public health research settings. The project will also serve as a proponent of improved recognition of data publications and other such published articles in formal academic reviews (recommendation 8, chapter **Error! Reference source not found.**) and other such settings.

Project tasks: Initially the project will involve engaging with the wider scientific community to determine the requirements of the portal. This will help me to begin building conceptual models of what the portal will look which I can then circulate for comments.

The next stage will be to build the underlying metadata model for the portal. Ideally, users will not see the metadata schema and will instead be presented with a user interface with a list of questions they will need to answer. For example, Simple Dublin Core maybe used to capture basic descriptive metadata. The second step would be to build or reuse an ontology to enable users to classify their metadata. Thirdly, a more comprehensive metadata standard such as DDI will be incorporated to capture lower level metadata.

Furthermore, as part of the user interface, advice on what constitutes good quality metadata will be provided as a means of supporting users. In doing so I will also be able to address in part recommendation statement 3: integrate metadata quality assessments into stakeholders' work routines as supported by increased provision of training and guidance.

The next stage is to run a pilot study to identify the strengths and weaknesses of the portal and determine how fit for purpose the registry is. Following these tests, the pilot portal will be extended to encompass all the envisioned functionality and made accessible globally. To encourage users to submit returns, this is something I will need to address with the input of the scientific community and the funder(s) involved. It may transpire that registration becomes a funder requirement and any unique identifiers assigned to a submission are to be included in any future publications along with details of where the data may be accessed / a request for access may be made.

Deliverables: The following lists the project deliverables (non-exhaustive)

- Report of findings from surveys conducted to engage with the wider scientific community and members of the public
- Novel metadata schema
- Ontology
- Pilot and finalised portals
- Provision of metadata quality assessment guidelines

Related successful grant application: This study builds on the work completed during the 'Enhancing Discoverability of Public Health and

Epidemiology Research Data' study findings from which were published in July 2014 (Castillo, Gregory et al. 2014). A total of 14 people from six organisations (four academic institutions and two from industry) were involved in delivering this project. My role in this study as project manager and responsibilities are detailed in chapter 1.5.4.

Related publications:

- **McMahon, C.** and S. Denaxas. (2016). "A novel framework for assessing metadata quality in epidemiological and public health research settings". *AMIA Summits on Translational Science Proceedings*. 2016;2016:199-208.
- **McMahon, C**., T. Castillo, et al. (2015). "Improving metadata quality assessment in public health and epidemiology." Stud Health Technol Inform **210**: 939.
- Castillo, T., A. Gregory, S. Moore, B. Hole, **C. McMahon**, S. Denaxas, V. Van den Eynden, H. L'Hours, L. Bell, J. Kneeshaw, M. Woollard, C. Kanjala, G. Knight, B. Zaba. (2014). "Enhancing Discoverability of Public Health and Epidemiology Research Data". London, United Kingdom.

Previous public engagement involving my project work: Posters were displayed at open day/evening events at UCL Institute of Child Health (Longitudinal studies and the Research Data Lifecycle: Application of the Data Documentation Initiative, 2012) and UCL Institute of Health Informatics (Improving metadata quality assessment in public health and epidemiology, 2015) to disseminate my findings and encourage discussion and debate around my work.