The benefits and challenges of linking health and administrative data with research data
A case-study review of using data linkage with longitudinal surveys

Margaret Kathryn Foley

A dissertation submitted to the University of Dublin,

in partial fulfilment of the requirements for the degree of

Master of Science in Health Informatics

2017

Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work, and has not been submitted as an exercise for a degree at this or any other university.

Signed: _______________________________ Date: ______________________

Margaret Foley

Permission to lend and/or to copy

I agree that the Trinity College Library may lend or copy this dissertation upon request.

Signed: ________________________________ Date: ______________________

Margaret Foley

Acknowledgements

This dissertation would not have been possible without the support and assistance of various people, to whom I am eternally grateful and possibly forever indebted, particularly:

- My supervisors Ms Gaye Stephens & Dr Christine McGarrigle for their continuous advice and guidance throughout this project
- Dr Lucy Hederman, Course Director, and all the lecturers and guest speakers on the course for their enthusiastic and dedicated teaching during the course
- My Health Informatics classmates – thank you for making this course such an enjoyable experience, for sharing not only your immense knowledge but also your friendship. Michelle and Dan thank you for keeping me motivated, laughing and full of coffee during our many weekends in the library
- My colleagues for your support and vital feedback on projects throughout this course, particularly during this dissertation, especially Siobhan Scarlett for her brilliant proofreading
- Cathal McCrory, Patrick Moore and Amanda Quail for their assistance with developing the questionnaire content
- All the longitudinal researchers who generously took the time to complete the questionnaire
- My family and friends, especially my parents and siblings, for putting up with all the “can’t talk, I’m in the library” responses to missed calls. Also, my brother Damien for his proofreading and the addition of many of the commas contained in this dissertation!

Abstract

There is a rising awareness of the significant potential for research to maximise the use and output of administrative data. Ireland currently lags behind other counties, many of which have established infrastructure for facilitating the use of administrative data in research projects for public benefit. The aim of this research was to identify the benefits and challenges of health and administrative data linkage with research data and explore how it could be facilitated in Ireland. Longitudinal research studies were selected as an appropriate example through which to examine these issues in the Irish context, as they have been the focus of linkage efforts in other counties with established linkage projects.

Relevant literature was reviewed to identity potential benefits and challenges to administrative data linkage. Additionally, primary research was conducted to collect both quantitative and qualitative data through a survey of longitudinal researchers, and by completing a privacy impact assessment on a sample linkage project.

Ultimately, the evidence shows that there are strengths and weaknesses of both data sources and that it is the combination of the two that enables research that otherwise may be impossible to achieve. While the responses to the questionnaire demonstrated that there are only a small number of linkage project ongoing in Ireland, there is an awareness of the potential benefits among longitudinal researchers to incorporating administrative data into their research. However, researchers were also aware that there are practical, cultural, legal and ethical challenges and implications which need to be addressed for the benefits of data linkage to be fully realised. Prominent challenges to emerge from this research are obtaining appropriate informed consent, the current lack of a unique identifier across administrative datasets and a reluctance of administrate data custodians to engage in linkage projects. The research also indicated that baseline knowledge of key issues such relevant legislation and regulations among researchers is relatively low which will impact any future linkage projects.

A prevailing demand from longitudinal researchers in Ireland to incorporating administrative data into their research analysis was identified. The evidence from this research supports the establishment of a national infrastructure to facilitate data linkage in Ireland, which would ensure any linkage is overseen by a national governance system. Establishing a national infrastructure would reduce the pressure on local resources to conduct individual data linkage and allow linkage expertise and matching technologies to develop within the national infrastructure.

Table of Contents

Appendixes

List of Figures

List of Tables

Abbreviations

ADRN	Administrative Data Research Network
ADT	Administrative Data Taskforce
ALSPAC	Avon Longitudinal Study of Parents and Children
ALSWH	Australian Longitudinal Study on Women's Health
AMIA	American Medical Informatics Association
BHPS	British household Panel Survey
CAG	Confidentiality Advisory Group
CNIL	National Commission on Informatics and Liberty (French Data Protection Office)
CSO	Central Statistics Office
DASSL	Data Access, Storage, Sharing and Linkage
DOB	Date of birth
DOHC	Department of Health and Children
DPC	Data Protection Commissioner
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
ED	Electoral District
ELSA	English longitudinal Study of Ageing
ERB	Ethics Review Body
ESRC	Economic and Social Research Council
EU	European Union
FaHCSIA	Department of Families, Housing, Community Services and Indigenous Affairs
GDPR	General Data Protection Regulation
GMS	General Medical Scheme
GP	General Practitioner
GUI	Growing up in Ireland
EHR	Electronic Health Record
HIMS	Health in Men Study
HIPE	Hospital Inpatient Enquiry system,
HIQA	Health Information and Quality Authority

HPO	Healthcare Pricing Office
HPSC	Health Protection Surveillance Centre
HRB	Heath Research Board
HRS	Health and Retirement Study
HSE	Health Service Executive
ICO	Information Commissioner's Office
IDS-TILDA	Intellectual Disability Supplement to The Irish Longitudinal Study on Ageing
IHI	Individual Health Identifier
ISSDA	Irish Social Sciences Data Archive
LDS	Longitudinal Data Set
LSIC	Longitudinal Survey of Immigrants to Canada
MAMMI	Maternal Health and Maternal Morbidity in Ireland
MCS	Millennium Cohort Study
NHS	National Health Service
NIDD	National Intellectual Disability Database
NSB	National Statistics Board
OECD	Organisation for Economic Co-operation and Development
PAC	Privacy Advisory Committee
PCRS	Primary Care Reimbursement Service
PI	Principle Investigator
PIA	Privacy Impact Assessment
PPS	Personal Public Service
RDT	Research Data Trust
RR	Response rate
SAIL	Secure Anonymised Information Linkage
SLID	Survey of Labour and Income Dynamics
SNOMED	Systematised Nomenclature of Medicine
TILDA	The Irish Longitudinal Study on Ageing
UI	Unique Identifier
WADLS	Western Australian Data Linkage System

1. Chapter 1: Introduction

1.1. Background and Motivation

The purpose of this research is to identify the benefits and challenges of health and administrative data linkage with research data and explore how it could be facilitated in Ireland. The research has been motivated by recent and impending changes in data protection legislation, health data standards and the development of eHealth Ireland, which will affect the potential to carry out administrative data linkage for research purposes in Ireland.

Administrative data is defined as information collected and used as part of the routine day-to-day provision or management of public sector services and schemes (MacFeely and Dunne, 2014). Health data is a subset of administrative data which has been collected in the course of providing healthcare. Throughout this project the term administrative data will be used to encompass both health data and wider administrative information. Utilising this data though data linkage involves bringing together, from two or more different sources, information that relates to the same individual, family, household or place (Holman et al., 2008, Leonard et al., 2013).

While the concept of data linkage in not a contemporary notion (Dunn, 1946), recent advancement in technology and health informatics has dramatically increased the possibilities in both data collection and linkage. Data linkage in other countries is enabled through the presence of a unique identifier across datasets, such as the National Health Service (NHS) number in the United Kingdom, which is key to ensuring effective matching (Hockley et al., 2008). Another factor which facilitates potential linkage projects is the increasing computerisation of administrative records (Tate et al., 2006, Audrey et al., 2016a). As highlighted by Calderwood and Lessof (2009), an increase in linkage is experienced when technical solutions such as these become available. As individual health identifiers (IHI) have recently been introduced in Ireland and electronic health records (EHR) are currently under development (HSE, 2017b, HSE, 2017a), it is a key time to review the potential for incorporating administrative data sources into Irish research.

This project has been precipitated by key pieces of work from national bodies such as the Health Research Board (HRB) (Moran, 2016) and the Health and Quality Authority (HIQA) (HIQA, 2012, HIQA, 2017c) as well as the impending adoption of the European General Data Protection Regulations (GDPR) (European Commission, 2016). These new regulations will represent a shifting landscape for the use of data in research.

In Ireland, there has recently also been an increase in resources dedicated to the exploration and development of systems that support the safe and secure sharing and linkage of data though initiatives endorsed, for example, by the National Statistics Board (NSB). However, while there has been a move to foster data sharing within and across government departments and agencies, such as the Department of Education and Skills using data linkage to track the education and economic status of school leavers (Tickner, 2013), there has been limited sharing of individual-level data outside the government sphere, restricting its incorporation into academic research (NSB, 2011). This is despite a rising awareness of the significant potential for research to maximise the use and output of administrative data as they contain a wide range of information of interest, not only to the policy research of government departments, but also the wider scientific community (Jones and Elias, 2006). Currently, Ireland lags behind other European countries, many of which have established infrastructure for facilitating the use of administrative data in research projects for public benefit, such as the Administrative Data Research Network (ADRN) in the United Kingdom (Boyd et al., 2014, ADRN, 2017).

For this research, longitudinal surveys were identified as a case study for reviewing the potential benefits and challenges of data linkage for research in Ireland as they have been the focus of linkage efforts in counties with established linkage projects such as Scotland and Australia (Jenkins et al., 2008, Hagger-Johnson, 2015). Additionally, the use of existing routinely collected data has been applauded as a method for enriching longitudinal surveys (Jones and Elias, 2006, Brett and Deary, 2014). While this research will focus on the implication of data linkage for longitudinal research projects, it is foreseen that the conclusions drawn could be applied to enriching a wider range of research and other secondary data uses, such as public health, policy development and audit, and evaluation of services.

1.2. Research Question

The research question *what are the benefits and challenges of linking health and administrative data with research data in Ireland* will be addressed by reviewing the existing international literature and exploring the current research environment for linkage with longitudinal research data in Ireland.

1.2.1. Aims and objectives of the research

The overall aim of the study is to identify "real world" benefits and challenges affecting researchers attempting to incorporate routinely collected health and administrative data sources into their research.

It is proposed that this will be achieved by:

1. Completing a review of existing literature of the benefits and challenges of data linkage with a focus on those encountered in the longitudinal research environment.

2. Conducting a survey of researchers working on Irish longitudinal studies to identify existing examples of data linkage being undertaken and assess the demand for further potential linkage projects.

3. Completing a privacy impact assessment (PIA) to identify the potential risks of a sample linkage project and determine if a PIA enables early identification of potential challenges.

4. Exploring the identified challenges, in combination with a review of the legislative and regulatory environment within which any future data linkage would occur, to assess how linkage can be facilitated through national infrastructure.

1.2.2. Overview of the research

Review relevant existing literature to identify benefits for, and challenges to data linkage in longitudinal research. This review will also explore key aspects of the proposal such as what constitutes data linkage and what are the errors or risks associated with the process.

Conduct a survey of longitudinal researchers to assess the demand for both linkage projects and linkage support and also to assess the baseline knowledge of researchers in relation to data protection, the Data Access, Storage, Sharing and Linkage (DASSL) model and HIQA standards. The survey questionnaire will include questions on any current linkage projects and how they were achieved, demand for future linkage projects and support, knowledge of linkage issues such as the DASSL model, and standards.

Complete a sample PIA of an administrative dataset linkage. The selected sample dataset will be identified from the survey of researchers and compare the identified risks against those identified during the literature review.

Review the evidence from the literature review, survey of researchers and the PIA to explore how administrative data linkage can be facilitated within the legislative and regulatory environment.

1.3. Overview of the Dissertation

The layout of the dissertations is as follows:

Chapter 1: Introduction – provides a backgrounds to the research subject matter and introduces the aims and objectives of the project.

Chapter 2: Literature Review – reviews relevant literature relating to the aims of the project, exploring topics such secondary use of data, fundamentals of longitudinal research and data linkage, and the benefits and challenges of data linkage.

Chapter 3: Relevant Legislation, Standards and Ethics – reviews the environment in which any future data linkage would occur by examining significant legislation, standards and ethics, which will be pertinent for the Irish research community.

Chapter 4: Research Methodology – outlines the approach utilised to answer the research question, the justification for the selected methods, an explanation of the collection tools and any ethical concerns.

Chapter 5: Results – presents the findings from both the survey of longitudinal researchers and the sample PIA, integrated with the evidence of benefits and challenges identified during the literature review.

Chapter 6: Discussion – discusses and interprets the results of this project, exploring the benefits and challenges of administrative data linkage and assessing how data linkage can be facilitated in Ireland through a proposed national infrastructure.

Chapter 7: Conclusion – summarises the findings of the project, addresses the limitations of the research and explores the implication for future practice and research.

2. Chapter 2: Literature Review

2.1. Introduction

As highlighted by the OECD in 2016, there is an immense amount of data collected for purposes other than research which, through data linkage, could be incorporated into research projects with the potential to vastly improve research capacity. However, the benefits of integrating these data sources into research must be balanced with any possible challenges or negative consequences for included research participants (Audrey et al., 2016a, OECD, 2016).

In order to fully explore the potential benefits and challenges of administrative data linkage, a comprehensive review of relevant literature will be completed. This review will also evaluate examples of existing administrative data linkage projects to assess how they achieved the "balance between the social value of research utilising such data and the protection of the well-being and rights, including privacy rights, of individuals" (OECD, 2016; pp.45).

Longitudinal surveys were identified as a case study to review these potential benefits and challenges of data linkage in Ireland as they have been the focus of linkage efforts in counties with established linkage projects such as Scotland and Australia (Hagger-Johnson, 2015). Additionally, incorporating administrative data into longitudinal studies was highlighted as one of the key recommendations by Martin et al. (2006) in their strategic review of ongoing studies in the UK.

2.2. Search Strategy

A comprehensive search of the available international literature in relation to data linkage was conducted with a focus on the potential impact for longitudinal research. This search included a wide view of the topic, including existing examples of linkage projects, the benefits of data linkage, the opportunities for future linkages and the challenges associated with any linkage projects. In addition to key databases, the publication lists of pertinent organisations and stakeholders were also searched, all of which are detailed in Table 2.1.

Table 2.1: Literature search sources

Databases	National Library of Medicine (NML) - Pubmed
	Cochrane Library
	Web of Science
	Scopus
Organisations/ Stakeholders	Department of Health & Children
	Economic and Social Research Council
	Economic and Social Research Institute
	Growing up in Ireland
	Health Information and Quality Authority
	Health Research Board
	Health Service Executive
	Intellectual Disability Supplement to The Irish Longitudinal Study on Ageing
	National Healthcare Quality Reporting System
	Organisation for Economic Co-operation and Development
	The Irish Longitudinal Study on Ageing

The main keyword search terms used to identify relevant literature during these searches were combinations of “data linkage”, “longitudinal” and “panel survey”. Searches were limited to articles published after 1997 to limit the results to the most recent material. While searches were not specially limited to the English language, any full text articles for which English translations could not be obtained were excluded from the literature analysis. Throughout the project the searches were repeated bimonthly to ensure any recent publications were included.

The titles and abstracts of the returned publications from these searches were reviewed for potential relevance and, if they were related to the research objectives, the full texts were obtained and reviewed. Additionally, the reference lists of the included articles were also reviewed to identify any further relevant literature that had been cited by the authors. While the review of the Cochrane Library returned no relevant papers, the results of the searches from the remaining databases are presented in Figure 2.1.

The following exclusion criteria were applied to the search strategy: while several articles were identified which detailed data linkage projects for research purposes, many of the studies were limited to linkage across administrative data sources, so were excluded from the literature analysis as they did not include a longitudinal study element (Lindgren et al., 2016, Kinnear et al., 2011). Also, some studies generated longitudinal data by linking administrative data from different time points but these were also excluded due to the lack of longitudinal survey data (Renzi et al., 2016, Hardelid et al., 2014, Eisenbach et al., 1997). These terms are explored further in section 2.4.

Additionally, some studies presented analysis resulting from successful administrative data linkage projects but did not address the potential challenges and benefits of the linkage process and were therefore excluded from a discussion around these issues. Details of all excluded studies are available in Appendix H and I.

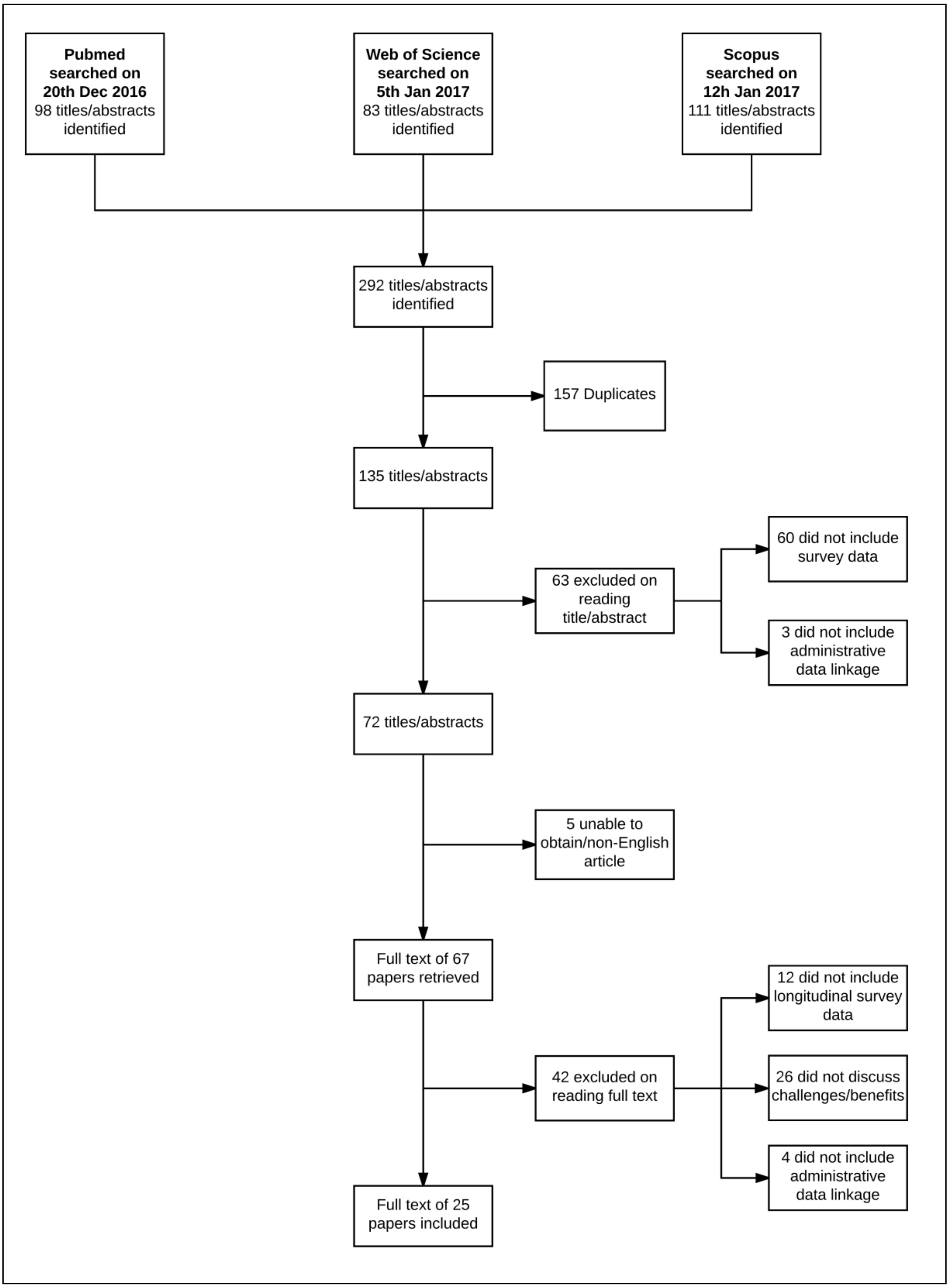

Pubmed searched on 20th Dec 2016
98 titles/abstracts identified
Web of Science searched on 5th Jan 2017
83 titles/abstracts identified
Scopus searched on 12h Jan 2017
111 titles/abstracts identified
292 titles/abstracts identified
157 Duplicates
135 titles/abstracts
63 excluded on reading title/abstract
60 did not include survey data
3 did not include administrative data linkage
72 titles/abstracts
5 unable to obtain/non-English article
Full text of 67 papers retrieved
42 excluded on reading full text
12 did not include longitudinal survey data
26 did not discuss challenges/benefits
4 did not include administrative data linkage
Full text of 25 papers included

Figure 2.1: Results of literature review search strategy of relevant databases

2.2.1. Development of themes for benefits and challenges

A thematic analysis strategy was applied to the review process (Guest et al., 2012). As the included papers were reviewed, reported benefits and challenges of data linkage began to emerge. In order to assess the implications of these, each issue reported in the papers was recorded in a database and coded into subthemes. Issues were identified in the literature were either added to an existing subtheme or new subthemes were created to accommodate them. This process of adding subthemes was continued until saturation was reached and no additional subthemes were emerging from the literature.

Once a complete set of subthemes had been established, they were aggregated into overarching themes for both benefits and challenges of data linkage. Developing these broader themes was an iterative process with several reviews and restructuring of the themes to ensure there was no duplication or crossover. A summary of the identified themes and subthemes, along with the number of articles/papers citing each them, are presented for the benefits and challenges in Table 2.2 and Table 2.3, respectively.

Despite being identified as themes in the literature review, data protection, data management standards and other relevant regulations are addressed separately in Chapter 3. This is due to significant incoming changes in Ireland in relation to these issues and also to fully explore the legislative and regulatory environment in which any future data linkage projects will occur.

Table 2.2: Themes and subthemes of administrative data linkage benefits and their distribution in reviewed literature

	Theme	Subtheme	Number of papers/articles
Benefits	**Data correction**	Addressing item non-response bias	7
		Reducing measurement error	5
		Reducing recall bias	10
		Addressing panel conditioning	4
	Data enrichment	Enable research that would otherwise not be possible by enhancing data	13
		Supplementing with data from outside collection period	5
		Supplementing with current data	1
	Sample maintenance and minimising attrition	Assessing sample representativeness	4
		Sample tracing	2
		Addressing attrition bias	4
		Providing data on participants who are withdrawn/lost to follow up/deceased	4
	Reduce participant burden	Reduce amount of data collected directly from participants	8
		Avoiding sensitive/embarrassing questions	1
		Allowing focused interview questions	2
		Issues with using administrative data	6
	Reduced costs and effective use of existing resources	Lower collection costs compared to traditional survey methods	9
		Effective use of existing resources	2
		Increasing the length of follow-up period	2
		Costs associated with reusing administrative data are offset	2

Table 2.3: Themes and subthemes of administrative data linkage challenges and their distribution in reviewed literature

	Theme	Subtheme	Number of papers/articles
Challenges	Consent	Informed consent required to enable research	5
		Can introduce bias	9
		Can reduce sample size	2
		Can compound section bias from original study sampling	4
		Varying rates of consent to data linkage	4
		Conflicting results on what influences consent to linkage	3
		Difficulty obtaining consent for data linkage	2
		Issues with attempting to link without consent	6
	Unique identifiers	Need for unique identifiers across datasets	6
		Issues with depending on unique identifiers for matching	4
		Alternatives to unique identifiers	2
	Data ownership and the role of data custodians	Unwillingness to share administrative data	7
		Resource implications for data custodians	2
		Lengthy approval processes	3
	Quality and structure of administrative data	Data not collected for sharing/reuse	7
		Data not designed for research purposes	3
		Only covers a proportion of the population	4
		Data quality issues	5
	Privacy and trust	Priority issue for participants	6
		Need for balance between privacy and research	6
		Difficulties quantifying privacy risks	6
	Technology limitations	Lack of digitalised administrative records	3
		Need for better matching technologies	4

2.3. Distinction between Primary and Secondary Use of Data

The primary use of data involves information being used for the purpose for which it was initially collected. Primary use of administrative data is defined as information collected and used as part of the routine day-to-day provision or management of public sector services and schemes (MacFeely and Dunne, 2014). Most government departments maintain records in relation to the services they provide and of the interactions citizens have with these services. (Jones and Elias, 2006, Calderwood and Lessof, 2009, MacFeely and Dunne, 2014).

In the case of routinely collected health data, its primary purpose is “protecting, promoting, maintaining or meeting the physical and mental health needs of an individual” (DOHC, 2009). Routinely collected health data includes, for example, hospital admissions records, prescription records and national disease registries (HIQA, 2016). Additionally, numerous other data sources are generated as part of administrative service delivery such as key demographic data including births and deaths certification, educational participation and attainment from school and examination board records and employment, income and tax details from taxation records (Calderwood and Lessof, 2009, Brett and Deary, 2014). While research is not the primary motivation for collecting this data, it often has significant research potential if reuse is possible (Moran, 2016). As highlighted by HIQA (2017c), the uses and benefits of high-quality data must be maximised to justify investing time, effort and resources into producing them.

In a recent review promoting an integrated approach to health and social care data, HIQA recommended that the reuse of routinely collected data should be optimised for secondary purposes such as research (HIQA, 2014b). This is echoed in the guiding principles relating to health information which dictate that health data should be collected once and used many times (HIQA, 2013). The support for the secondary use of data is echoed outside of the health sphere with the Central Statistics Office (CSO) establishing a national data infrastructure to facilitate the integrated use of data collected across government departments (NSB, 2015).

The secondary use of data has been defined by the American Medical Informatics Association (AMIA) as the use of data beyond the purpose for which it was originally collected, such as “analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities” (Safran et al., 2007; pp.4). As highlighted by the Economic and Social Research Council (ESRC), administrative data such as those described above have the potential to provide a rich evidence base than can contribute to research as well as policy development and evaluation (ADT, 2012).

2.4. Fundamentals of Longitudinal Research

2.4.1. Longitudinal study design

Longitudinal research involves collecting information from the same individuals or households at several points in time. As illustrated in Figure 2.2, a key aspect of longitudinal studies is the repetition of questions and measures over multiple time points in order to assess cumulative effects and patterns of change over time (Rajulton, 2001, Lynn, 2009, Hagger-Johnson, 2015). Owing to the nature of the repeated data collection, they can be used to obtain better information about causal relationships and evaluate the cumulative effects of social, physical and environmental exposures on the human life course (Martin et al., 2006). Due to these advantages, longitudinal studies have been described as the "cornerstone" of social science research and the number of longitudinal studies being conducted has increased in recent years with growing interest from academia, government and private sectors (Lynn, 2009, Townsley, 2016).

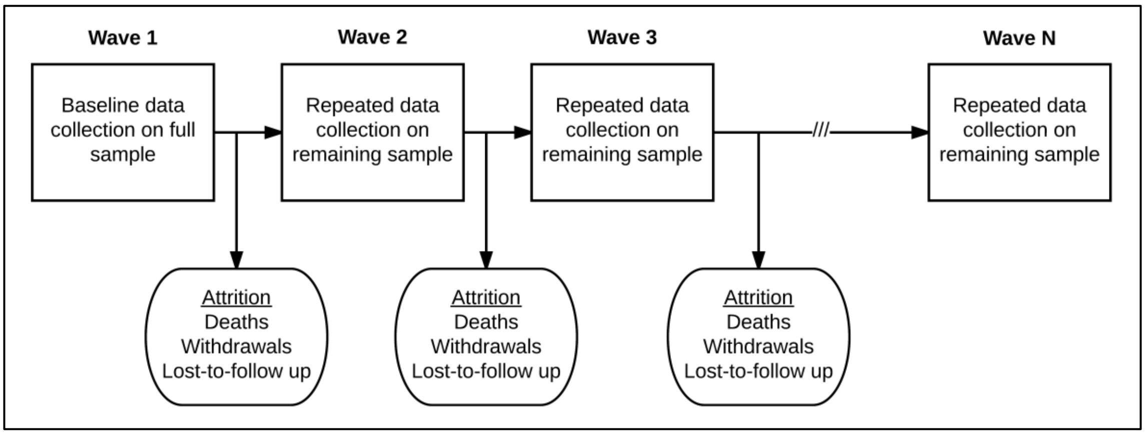

Wave 1
Wave 2
Wave 3
Wave N
Baseline data collection on full sample
Repeated data collection on remaining sample
Repeated data collection on remaining sample
///
Repeated data collection on remaining sample
Attrition
Deaths
Withdrawals
Lost-to-follow up
Attrition
Deaths
Withdrawals
Lost-to-follow up
Attrition
Deaths
Withdrawals
Lost-to-follow up

Figure 2.2: Longitudinal study design

Longitudinal studies are frequently identified as key resources for addressing wide-ranging topics, such as demographic shifts, cultural diversity, socioeconomic inequalities and ageing populations (Martin et al., 2006, UKDF, 2013, ESRC, 2015). The ESRC identified longitudinal studies as a flagship resource, both for investigating individual life-course development due to the unique ability to study the effects of earlier characteristics on later outcomes, and also for their value in addressing key scientific questions in relation to societal well-being and policy development and evaluation (Martin et al., 2006).

The focus of longitudinal studies covers a range of disciplines such as sociology, health and medicine, psychology, economics, politics, demographics and environmental science with many of the studies combining disciplines to achieve a full picture of the selected cohort and prevent unnecessary duplication or the need for multiple studies. The wide range of data collected on each individual enables the investigation of multiple covariates while controlling for confounders (Martin et al., 2006). Longitudinal studies can also facilitate 'natural experiments', such as evaluating policy changes, by comparing data collected before and after a change has been implemented (Townsley, 2016).

2.4.2. Types of longitudinal studies

There are several types of longitudinal survey designs (Table 2.4) with the defining feature usually how the original cohort are defined and selected. For example, some longitudinal studies select a cohort consisting of a set age range or birth period, whereas others focus on obtaining a defined cohort of individuals with similar features to investigate a specific population or condition. Longitudinal data can also be constructed entirely from administrative sources (Martin et al., 2006). However, overlapping and nesting can occur between the study designs such as completing an area study on a subsample of a larger household panel study (FaHCSIA, 2013).

Table 2.4: Longitudinal survey designs

Longitudinal Study Design	Features	Examples
Household/family panels	Chart family life and household change	British Household Panel Survey (BHPS)
Birth cohorts	Sample of individuals all born at the same time (same day, week, month or year)	Millennium Cohort Study (MCS)
Age cohorts	Sample of individuals all within a set age group or at a set transition such as school entry or completion	English Longitudinal Study of Ageing (ELSA)
Special population studies	Focus on small population groups such as ethnic minorities or immigrants	Longitudinal Survey of Immigrants to Canada (LSIC)
Area studies	Data collected from individuals or families and the local institutions and services to which they relate	Avon Longitudinal Study of Parents and Children (ALSPAC),
Record linkage studies	Relies entirely on the use of administrative records	Administrative Longitudinal Data Set (LDS) - Australia

Source: modified from Martin et al (2006)

Collected data may be quantitative or qualitative or a combination of both. The data collection methods can also vary across studies with methods such as face-to-face interviewing, telephone interviewing, postal questionnaires and physical examinations being utilised, with many studies using a mixed-mode approach to obtain a greater breadth and depth of data (Golding and Jones, 2009).

2.4.3. Examples of longitudinal studies

Established longitudinal studies are ongoing in countries across the world, such as the Health and Retirement Study (HRS) in the United States, the Understanding Society study in the United Kingdom and the Australian Longitudinal Study on Women's Health (ALSWH). Presently in Ireland, there are four established longitudinal surveys being conducted. These are detailed in Table 2.5 along with a brief summary of their focus, study type and international equivalents as many of the studies are developed within an international community.

However, there are other current international longitudinal studies which do not have Irish equivalents. For example, Australia has separate studies to investigate the health outcomes of each gender (Australian Longitudinal Study on Women's Health; The Australian Longitudinal Study on Male Health) and Canada has developed a Longitudinal Survey of Immigrants. Due to the smaller population in Ireland and limited resources, the number of studies that can be conducted is constrained and it is therefore vital to maximise the potential of the existing studies to effectively address a wide range of research topics.

Table 2.5: Examples of established longitudinal studies in Ireland

Study Name	Focus	Study Type	International equivalent studies
Growing up in Ireland (GUI)	Two cohorts of children aged 9 years (child cohort) and 9 months (infant cohort).	Age cohort	Growing up in New Zealand
The Irish Longitudinal Study on Ageing (TILDA)	8,500 community dwelling adults aged 50 and over	Age cohort	English Longitudinal Study of Ageing (ELSA)
The Intellectual Disability Supplement to TILDA (IDS-TILDA)	753 people with an intellectual disability aged 40 and over.	Special population study	Longitudinal Health & Intellectual Disability Study - USA
Maternal health And Maternal Morbidity in Ireland (MAMMI study)	Pregnant women having their first baby	Special population study	The Western Australian Pregnancy Cohort (Raine) Study

2.4.4. Limitations of longitudinal studies

Although they are highlighted as a valuable methodological design, it is vital to remember that longitudinal studies have inherent limitations (Rajulton, 2001). The study design depends heavily on the information provided directly from participants which may be subject to errors in self-reporting and can be severely affected by attrition which limits potential analysis (Brett and Deary, 2014). In order for longitudinal studies to continue to contribute to future research output, new and innovative ways to address the limitations described below are required.

2.4.4.1. Item non-response

A major methodological challenge for longitudinal surveys is missing data caused by item non-response (Lynn, 2009). This occurs when a participant agrees to partake in the study but fails to provide answers to individual questions within the study. The resulting missing data can have significant impact on data quality and statistical analysis, particularly if the missing data trends are associated with the variables of interest (Yan and Curtin, 2010, Mars et al., 2016).

2.4.4.2. Measurement error

Measurement error is defined as the possibility that a "survey observation might differ from the value that would be observed by a perfect measurement" (Lynn, 2009; pp.16). Essentially, it is an error introduced to the data due to ineffective collection methods resulting in collected data which is not a true reflection of reality. It can be introduced through a number of different channels including the measurement tool, the collection mode (face-to-face interview, self-completion, etc.), the participant, the interviewer or the interview setting (Biemer and Lyberg, 2003). While this methodological issue is not limited to longitudinal data collection, Martin et al. (2006) identified measurement error as a major threat to the quality of longitudinal studies as it is magnified because much of the analysis involves the measurement of change over time which is highly sensitive to measurement error (Martin et al., 2006, Lynn, 2009).

One type of measurement error commonly encountered is recall error, which is caused by inaccuracy or incompleteness in how participants remember or report on past events or experiences (International Epidemiological Association, 2014). Recall error can occur in several formats such as an event being completely forgotten, the timing of an event being remembered incorrectly or any associated causes or consequences being misremembered (Lynn, 2009). For example, a participant may be able to accurately recall if and when they had a heart attack, but they may find it much more difficult to correctly remember how many times they visited their General Practitioner (GP) in the past two years. This reduced accuracy in recalling events that are neither rare, salient nor recent has the potential to introduce bias into a study and negatively impact analysis results (Tourangeau et al., 2000). In the example above, any policy developed using the reported GP visits may be incorrectly informed and as a result not effectively address GP utilisation for the population.

2.4.4.3. Panel conditioning

Another key limitation is panel conditioning, which is unique to longitudinal studies. As participants complete repeated interviews, they can become familiar with the dynamic routing which leads to additional questions depending on how a leading question is answered. For instance, reporting a chronic health condition would lead to several follow-up questions to collect further information about the condition and how it affects the participant. If a participant wishes to avoid questions or shorten their interview they may deny the presence of any conditions leading to inaccurate data collection (Rajulton, 2001, Martin et al., 2006, Halpern-Manners et al., 2014).

2.4.4.4. Sample attrition

One of the major concerns for longitudinal researchers is maintaining the original study sample through participant retention. Sample attrition is defined as “the continued loss of respondents from the sample due to nonresponse at each wave of a longitudinal survey” (Lynn, 2009; pp.10). Attrition is often considered to be the ‘Achilles Heel’ of longitudinal surveys and can occur for several reasons; some participants may choose to withdraw while others are lost-to-follow-up and cannot be traced at subsequent waves. Additionally, surveys will experience attrition due to participants passing away, which is a greater concern in studies of older populations (Martin et al., 2006).

Accumulated attrition over consecutive waves of data collection, can result in a sample size that is no longer representative of the population from which it was drawn or unacceptably small, limiting high quality empirical research, which may affect the feasibility of continuing with data collection (Martin et al., 2006). Additionally, attrition from the sample is rarely random and so can result in bias being introduced to the analysis (Chatfield et al., 2005, Watson and Wooden, 2009, Eapen et al., 2014). As a result of these issues, minimising attrition is a major concern for longitudinal researchers and strategies to address the issue are a key focus for longitudinal methodological research (Martin et al., 2006, Lynn, 2009).

2.4.4.5. Costs of data collection

Prospective longitudinal studies are extremely resource heavy, requiring large amounts of initial outlay to develop and establish them, coupled with a need for long-term investment to ensure they are maintained for a period sufficient enough to collect meaningful longitudinal data (Martin et al., 2006).

2.5. Data Linkage

Data linkage has been described as the bringing together, from two or more different sources, information that relates to the same individual, family, household or place (Holman et al., 2008, Leonard et al., 2013). Linkage projects are usually conducted on large population-based research or projects in order to maximise the output (Mountain et al., 2016).

The methodology of data linkage requires comparing key shared variables from the records of each separate dataset in order to determine and join records that relate to the same entity (Churches et al., 2002). When linking records for individuals, common matching variables can include administrative identifiers, name, gender or date of birth (DOB), whereas attempting to link households or places may rely on postcodes or geocodes as the matching variables (Jones and Elias, 2006). Data linkage aims to maximise the number of correctly linked records from across the different datasets, however the process must be conducted in a precise manner in order to ensure records are not incorrectly matched which can undermine any resulting analysis of the data.

Internationally, and to a smaller extent in Ireland, linkage projects between longitudinal studies and administrative datasets are already occurring. For example, in Ireland, TILDA has been linked with prescribing records for 72% of the participants who had a medical card and consented to the linkage (Moriarty et al., 2015). However, there is significant variability between countries in terms of the availability and accessibility of administrative data for linkage with research data (Martin et al., 2006). Scandinavian countries, for example, have a strong culture of data linkage and an established system of assigning unique identifiers at birth. Other countries such as Australia and Scotland also have established collaboration between administrative data and research studies.

For example, in Australia the Western Australia Data Linkage System (WADLS), which was established in 1995, contains data from over 30 administrative collections since 1966 with monthly updates to maintain current data (Holman et al., 2008). De-identified data is made available for researchers and has been linked to several longitudinal studies including ALSWH, the Western Australian Pregnancy Cohort Study, the Health in Men Study (HIMS) and the Fremantle diabetes study (Mountain et al., 2016, Yeap et al., 2013, Almeida et al., 2012, Tooth et al., 2012, Hart et al., 2015).

2.5.1. Data linkage errors

Data linkage can be subject to two types of linkage errors (Table 2.6). Firstly, false matches where the linkage process incorrectly matches two records that do not actually relate to the same person, referred to as false-positive errors. Secondly, missed matches where the process fails to recognise that records relate to the same person, which are termed false-negative errors (Hagger-Johnson et al., 2015). Attempts to reduce one of these types of errors tends to lead to an increase in the other, and as a result a common approach is to attempt to minimise the sum of the two error types (Calderwood and Lessof, 2009). Assessing the likelihood of these errors occurring is essential when choosing an appropriate data linkage method as this will determine the potential level of error in the final dataset.

Table 2.6: Classification of data linkage errors

Outcome of linkage	Records relate to the same person	Records relate to different persons
Linked records	True positive	***False positive***
Unlinked records	***False negative***	True negative

Source: modified from Calderwood & Lessof, 2009

2.5.2. Data Linkage Methods

There are varying methods of matching individuals across different datasets. Matching can be deterministic, where an exact match on all linking variables is required. Deterministic methods are often used where high quality unique identifiers are available across the datasets. However, it is also possible to carry out exact matching using combinations of variables such as name, address, gender and DOB. Using names or addresses for exact matching can be problematic as they are rarely unique and are often subject to variation in recording and spelling. For example, the use of nicknames or truncation in one dataset may cause false-negatives when linkage is attempted (Gill, 2001, Jenkins et al., 2008). As highlighted by Calderwood and Lessof (2009), due to clerical or transcribing errors, even with the availability of a unique identifier, depending on deterministic matching can results in false positives or false negatives.

Probabilistic matching, which utilises non-exact matching, can be attempted in these case where exact matches would not be feasible. Probabilistic matching can be favoured over deterministic methods as all data is subject to potential error. With this method, a set tolerance level of mismatch between two records is allowed. For example, a potential match where the DOBs in the two datasets differs by one digit would be allow as long as all other linking variables have an exact match. The matching is determined by the probability that identified matches and non-matches are true or false (Gill, 2001, Martin et al., 2006). Essentially, this method accepts near-matches as relating to the same individual and determines whether the records should be linked based on an agreed margin of error. The potential for false positive and negative matching errors occurring can be controlled by adjusting the accepted near-match tolerance levels that determine a link. Lowering the tolerance level, for example by permitting DOB difference within one year rather than one month, may reduce the amount of false negative matches that occur but is likely to increase the level of false positive errors (Jones and Elias, 2006).

2.6. Benefits of Administrative Data Linkage

A review of the literature revealed many reported benefits of linkage projects. Interestingly, these benefits appear to be bidirectional with positive outcomes reported for both individual studies and the wider research community but also for the custodians of the administrative data (Jones and Elias, 2006, Calderwood and Lessof, 2009). These benefits, as presented in Table 2.7, will be explored under the themes and subthemes developed during the thematic analysis detailed in Section 2.2.1.

Table 2.7: Benefits of administrative data linkage

Benefit	Subtheme
Data correction	Addressing item non-response bias
	Reducing measurement error
	Reducing recall bias
	Addressing panel conditioning
Data enrichment	Enable research that would otherwise not be possible by enhancing data
	Supplementing with data from outside collection period
	Adding current/continuous data
Sample maintenance and minimising attrition	Assessing sample representativeness
	Sample tracing
	Addressing attrition bias
	Providing data on participants who are withdrawn/lost to follow up/deceased
Reduce participant burden	Reduce amount of data collected directly from participants
	Avoiding sensitive/embarrassing questions
	Allowing focused interview questions
	Issues with using administrative data
Reduced costs and effective use of existing resources	Lower collection costs compared to traditional survey methods
	Effective use of existing resources
	Increasing the length of follow-up period
	Costs associated with reusing administrative data are offset

2.6.1. Data correction

Longitudinal studies are uniquely placed to address complex research questions. Research studies are heavily dependent on the quality of their data to produce prominent scientific output and ensure return to investment. Yet due to their design, longitudinal studies are subject to inherent methodological issues, as detailed in Section 2.4.4, which can affect their ability to address research questions. However, administrative data linkage presents a potential solution to some of these methodological weaknesses.

2.6.1.1. Addressing item non-response

As highlighted in Section 2.4.4.1, item non-response represents a methodological challenge for longitudinal surveys which can have a significant impact on data quality and statistical analysis (Lynn, 2009). Supplementing survey data with information from administrative data sources was proposed as a solution to address missing data within surveys (Martin et al., 2006, Jones and Elias, 2006, Cornish et al., 2015, Audrey et al., 2016a, Audrey et al., 2016b, Mountain et al., 2016, Mars et al., 2016). This is particularly important if the missing data trends are associated with the variables of interest as it can impact on the analysis of the study. In one example, while investigating adolescent self-harm, Mars et al. (2016) used administrative data linkage to show the prevalence of self-harm was higher in participants who did not respond to the self-harm questions. Similarly, administrative data linkage can be used to correct underreporting, undervaluing or rounding by participants which is often encountered when collecting information on income and assets (Jones and Elias, 2006, Pudney, 2008, Mars et al., 2016).

2.6.1.2. Reducing measurement error

One of the main reported benefits of supplementing research data with administrative data is the potential to minimise measurement error (Jenkins et al., 2008). As detailed in Section 2.4.4.2, measurement error is a major threat to the quality of longitudinal studies as its effect is magnified in the repeated measurements over time (Martin et al., 2006, Lynn, 2009). By linking with administrative data records, researchers can assess the quality and accuracy of the data obtained in the survey and correct inaccurate data where possible, improving the data quality (Holman et al., 2008, Sala et al., 2012, Brett and Deary, 2014, Al Baghal, 2016, Audrey et al., 2016a, Mars et al., 2016). Such comparison studies facilitate validation of self-reported data against administrative records which can improve the overall quality of the research (Mountain et al., 2016). For example, a data validation study of participants' father's occupation reported by Rajulton (2001) found only a 70% agreement between the information provided by the participant and the data contained in census records.

Administrative data linkage is particularly well placed to address the issues associated with recall bias, as detailed in Section 2.4.4.2. Due to nature of administrative data collection, it is less likely to be subject to recall error and therefore, in some cases, is more accurate than data obtained during the survey (Jones and Elias, 2006, Calderwood and Lessof, 2009, Husain et al., 2012, Knies et al., 2012, Hure et al., 2015). Returning to the example of GP visits, these events would be recorded in health records as they occur and consequently no recall is required. Similarly, information on income and earnings obtained from tax records would be based on the amounts reported to Revenue by the employer each year rather than based on participant recall.

However, it is important to acknowledge that measurement error and missing data can also occur within administrative data, and Martin et al. (2006) cautions against overlooking the risk of error in administrative datasets when linking with research data. Additionally, the issue with recall bias was not replicated in all the identified literature, with Carroll et al. (2016) reporting overall concordance of healthcare utilisation data obtained from self-report compared to administrative records. However, the recall periods examined in this study were only up to six months, with the highest agreement reported after just one month, and therefore were relatively short when compared to other longitudinal studies' recall periods (Martin et al., 2006).

2.6.1.3. Addressing panel conditioning

An additional methodological issue which has the potential to be addressed by incorporating administrative data is panel conditioning and question avoidance. Rajulton (2001) highlights that the literature exploring this form of bias is sparse, as fully investigating its effects can require costly resources such as a longitudinal control groups. Combining research and administrative data may provide a solution to both investigating the occurrence of conditioning and also correcting the collected data by acting as a comparison and therefore can inform and improve survey methodology (Al Baghal, 2016). For example, participants who report taking chronic medications in a baseline wave but none are reported at subsequent waves could be compared against medication reimbursement datasets to ascertain if the medication is still being prescribed.

2.6.2. Data enrichment

In addition to data correction, linkage with administrative sources can be used to enrich and enhance the data collected through longitudinal surveys.

2.6.2.1. Enable research that would otherwise not be possible by enhancing data

Administrative data linkage can enable greater opportunities for research by widening the evidence base by supplementing survey data with additional information that would be difficult or impossible to collect through traditional survey methods (Jones and Elias, 2006, Martin et al., 2006, Jenkins et al., 2008, Fredman et al., 2001, Sakshaug et al., 2013, Grusky et al., 2014, Hagger-Johnson, 2015, Calderwood and Lessof, 2009, Brett and Deary, 2014, Audrey et al., 2016a, Mostafa, 2016).

One example of data enrichment through linkage is the addition of small area statistics, such as at electoral district level, from the area in which a participant lives. This can enable research on local facilities and exposures by providing information that would not be available directly from the participant, such as access to green spaces and healthcare services or levels of radon and noise pollution (Martin et al., 2006).

2.6.2.2. Supplementing with data from outside collection period

Administrative data linkage can also be used to supplement research data with information that relates to events which occur outside of the study collection period, such as birth or mortality data (Fredman et al., 2001, Martin et al., 2006, Calderwood and Lessof, 2009, Husain et al., 2012). Mountain *et al.* (2016) also recommends utilising data linkage for retrospective collection of measures not included in earlier waves of a study in order to increase follow-up periods or to investigate emerging theories and associations. This benefit was echoed throughout the literature with several papers reporting data linkage facilitated addressing research questions which were previously unanswerable and in turn accelerated research as existing data could be exploited rather than requiring new studies to investigate emerging concepts (Jones and Elias, 2006, Soloff et al., 2007, Holman et al., 2008, Calderwood and Lessof, 2009, Sakshaug et al., 2013, Grusky et al., 2014, Hagger-Johnson, 2015, Townsley, 2016).

2.6.2.3. Supplementing with current data

While there are benefits of using data linkage to incorporate historical or retrospective data as described above, Jones and Elias (2006) also emphasise that it can be used to supplement survey data with current information. As administrative records are usually regularly or continuously updated, data from recent time periods can be incorporated into analysis without the need to wait for the next wave of surveying. ELSA employ this method to obtain up-to-date information on cancer and mortality from the Health and Social Care Information Centre. Once a participant has consented, a flag is added to their Information Centre record and if they are diagnosed with cancer or pass away, ELSA is notified with cancer details or causes of death which can be linked back to the participant's survey data (ELSA, 2015).

2.6.3. Sample maintenance and minimising attrition

Many of the studies identified from the literature review endorsed data linkage as a method for addressing sample maintenance and participant attrition.

2.6.3.1. Assessing sample representativeness

Longitudinal studies aim to be representative of the population from which the sample is drawn. Taking the baseline sample selection for TILDA as an example, it ensured that the participants are representative of the Irish population aged 50 years or older and means any results extrapolated from the study can be applied to the whole population (Whelan and Savva, 2013). As a longitudinal study continues through repeated waves of collection, linkage with overall population statistics allows ongoing comparison of the study cohort against the population from which it is drawn. This can facilitate assessment of the study's representativeness and can highlight subpopulations which are underrepresented or need to be focused on during any replenishment of the sample (Eapen et al., 2014, Husain et al., 2012, Mountain et al., 2016).

Administrative databases can also be integrated in the design of a survey and used to develop the original sampling frame, meaning it can be used to establish effective weighting strategies and establish non-response bias at baseline (Calderwood and Lessof, 2009, Macleod et al., 2010, Hagger-Johnson, 2015).

2.6.3.2. Sample tracing

Maintaining the sample over the lifetime of a study is dependent on successful tracking and contacting participants at each wave of collection (Lynn, 2009). Administrative records have a distinct advantage over surveys when tracing participants as people are more likely to remain engaged with the services linked to the administrative records and keep them updated with current contact details (Jones and Elias, 2006, Calderwood and Lessof, 2009). Growing Up in Ireland (GUI), for example, seeks consent to link with the Child Benefit Register operated by the Department of Social and Family Affairs in order to trace participants who have moved between waves or cannot be traced by the interviewers (Thornton et al., 2013). This method of participant tracing was also employed by Brett and Deary (2014) in their efforts to re-contact all participants from a historical longitudinal sample using the National Health Service Central Register.

2.6.3.3. Providing data on participants who are withdrawn/lost to follow-up/deceased

As described in Section 2.4.4.4, sample attrition is perceived as the 'Achilles' heel' of longitudinal research. Examples from the literature show that authors used data linkage to continue to collect data on participants who had been lost to follow-up in order to reduce the effects of attrition bias (Jones and Elias, 2006, Lessof, 2009, Mountain et al., 2016). This is particularly important when investigating conditions which influence attrition, such as dementia in ageing cohorts. As demonstrated by Chatfield et al*.* (2005), cognitive decline in older populations will be underestimated if the attrition bias, which is seen consistently across studies, is not adjusted for accordingly. Hagger-Johnson (2015) used consent for data linkage, obtained during earlier waves, to continue to follow-up on the health outcomes of participants who were no longer taking part in prospective interviews, meaning they could continue to be included in longitudinal analysis. This method is also used to determine if participants who haven withdrawn or been lost to follow-up are deceased through linkage with death registries (Brett and Deary, 2014).

2.6.3.4. Addressing attrition bias

In addition to being used to collect follow-up information on participants who attrit, data linkage can provide data on the characteristics influencing attrition and, as a result, analysis can be more accurately adjusted to account for attrition bias (Watson and Wooden, 2009, Eapen et al., 2014, McGhee et al., 2015). Longitudinal studies are subject to healthy survivor effects, where the unhealthier participants are more likely to attrite, but using linkage with health records, researchers can compare the characteristics of those who continue with the study with those who don't, quantifying the extent of health selection in the longitudinal sample (Hagger-Johnson, 2015).

2.6.4. Reduce participant burden

Due to the consequences of attrition, as outlined in Section 2.4.4.4, it is vital that participants in longitudinal research find the survey a pleasant experience as those who, for example, find it too long, difficult, embarrassing or uninteresting, are less likely to take part at subsequent waves (Lynn, 2009). Therefore, efforts must be made to reduce the burden placed on participants to secure their continued involvement in the study, and evidence from the literature suggested administrative data linkage can assist with reducing this burden.

2.6.4.1. Reduce the amount of data collected directly from participants

Many of the collection methods used in longitudinal surveys, such as interviews, self-completion questionnaires, diaries, physical assessments and biological sampling, generally involve considerable time and commitment from the participants (Mountain et al., 2016). However, as indicted by Martin et al. (2006) many of the details collected during the survey, such as income and tax details, are time consuming and tedious to collect and are already available in administrative databases. Therefore, by using data already available through administrative data sources, researchers can reduce the amount of information that needs to be collected directly from the participants (Soloff et al., 2007, Jenkins et al., 2008, ADT, 2012, Sala et al., 2012, Al Baghal, 2016, Audrey et al., 2016a, Mountain et al., 2016).

The Canadian Survey of Labour and Income Dynamics (SLID), for example, offered participants the option of answering 25 questions in relation to income directly or alternatively allowing the data to be collected through linkage with tax records (Michaud et al., 1995). While the authors reported a positive impact on participant burden, other studies cautioned that these mixed-methods may introduce the errors of both types of data if the two sources are not used to supplement each other correctly, due to the varying data structures and quality across the two sources (Calderwood and Lessof, 2009). Additionally, Sakshaug (2013) stresses that making participants aware that their burden will be reduced, through a shorter interview, does not appear to impact on their propensity to consent to data linkage.

2.6.4.2. Allowing focused interview questions

Alternatively, Calderwood and Lessof (2009) propose, rather than promoting a shorter interview, researchers use the time saved by removing questions on data that can be obtained through linkage to focus on more interesting topics or to collect information that would not be available through administrative sources, such as personality indicators or self-rated measures of health and well-being. In this way, the strengths of the two data sources are maximised with administrative data be utilised for routine objective information or details that are prone to measurement error when collected directly from participants and survey collection used to collect more subjective details or information which is not usually incorporated into administrative data.

2.6.4.3. Avoiding sensitive/embarrassing questions

As suggested by Calderwood and Lessof (2009), data linkage can also be used to avoid asking participants directly about subjects that may be sensitive or embarrassing such as addiction, literacy issues or whether they are experiencing incontinence. Collecting this data directly from participants may make them uncomfortable to such an extent that they no longer wish to participate in the study or they may be too embarrassed to admit to such issues and so provide inaccurate data.

2.6.4.4. Issues with using administrative data to reduce participant burden

Evidence from the literature also cautions against depending solely on administrative data to detect these sensitive issues as in many cases the participants may not have informed their healthcare provider either. Taking the example of incontinence, recent research from TILDA suggest that of those participants who reported experiencing incontinence to the survey interviewer, only 3 in 5 had informed a healthcare provider (Canney et al., 2016). As a result, depending on the information from administrative records would significantly underestimate this condition and have serious implications for any analysis drawn from the data (Knies et al., 2012). These concerns were also echoed for financial data, with the authors of the SLID data linkage expressing concerns that income from ‘underground economies’ would be missed by depending solely on the data from administrative sources (Michaud et al., 1995). Therefore, Mars et al. (2016) recommended a combination of both survey and administrative data to obtain accurate figures (Knies et al., 2012, Mars et al., 2016).

2.6.5. Reduced costs and effective use of existing resources

The initial and ongoing costs of a longitudinal study are inherently high due to the long-term nature of the studies but also the extensive planning that is required to ensure the chosen sampling model effectively represents the population and that the study design and collection methods are appropriate to achieve the aims of the study (Lynn, 2009, Brett and Deary, 2014). Several of the papers identified in this review, reported potential cost mitigation as an incentive to implementing data linkage.

2.6.5.1. Lower collection costs compared to traditional survey methods

As demonstrated by Calderwood and Lessof (2009), obtaining data from existing administrative data sources is relatively faster and cheaper when compared to the high costs and resources required for survey collection and, as more administrative records are digitalised, it will become even easier to generate new research using existing data sources rather than generating new primary data. (Jones and Elias, 2006, Jenkins et al., 2008, ADT, 2012, Sala et al., 2012, Sakshaug et al., 2013, McGhee et al., 2015, Al Baghal, 2016, Audrey et al., 2016b).

2.6.5.2. Effective use of existing resources

Part of the wider cost reduction associated with data linkage is through minimising the collection of duplicate data, as surveys often replicate data that are already collected as part of routine administrative data (Jones and Elias, 2006). This reuse of existing data can therefore represent effective use of resources for the administrative data owners by utilising data that may otherwise remain dormant, resulting in cost-saving not only for the individual studies but also government departments, other stakeholders and the wider public (Jones and Elias, 2006).

2.6.5.3. Costs associated with reusing administrative data are offset

While there are costs associated with the extraction and preparation of administrative data, these are offset by the potential savings gained by reducing the amount of data collected through resource heavy surveying methods. Holman et al. (2008) highlighted that reusing and integrating administrative data in this way, generates a return on investment for the significant resources that are dedicated to developing and maintaining administrative datasets. This potential benefit, through secondary use of administrative data, could therefore be used to justify dedicating resources to data sources which can be incorporated into research, such as

EHRs. For example, as of 2008, the WADLS data linkage service in Australia had enabled administrative data to support over 400 research studies, resulting in over 250 journal publications and is also used to facilitate a national education programme on analysing linked heath data which represents a significant reuse of administrative data resources. Additionally, grant funding which was secured through the infrastructure, represented more than a tenfold return on investment (Holman et al., 2008). This benefit was echoed by Hagger-Johnson (2015), who reported an increase in grant funding due to the ability to address new research questions and implement novel research methodologies.

2.6.5.4. Increasing the length of follow-up period

Data linkage can also facilitate cost savings for researchers by extending a study's follow-up period, which may allow for more meaningful analysis, particularly in the case of life-course research (Audrey et al., 2016a). A review of current and historical longitudinal studies of ageing established that, while the range of follow-up period varied from two to over thirty years, for the majority of studies the follow-up was less than ten years which may limit their ability to investigate the long-term influences of ageing (Seematter-Bagnoud and Santos-Eggimann, 2006). One example presented by Brett and Deary (2014) demonstrated how data linkage could be utilised to extend the follow-up period without the costs associated with resurveying participants, by using a longitudinal study which collected data from early childhood during the 1940's and 50's and linkage to current hospital and death records to give a follow-up period which spanned the full life-course. One of the distinct advantages of extending the follow-up period as detailed above, is that it avoids any further burden directly for the survey participants (Jones and Elias, 2006, Calderwood and Lessof, 2009). These methods can also be of particular benefit to further exploit data from historical longitudinal surveys which are no longer actively collecting primary data.

However, while the cost savings presented above may suggest abandoning longitudinal survey collection completely in favour of the cheaper reuse of existing administrative data, it is vital to remember that there are strengths and weaknesses of both data sources and that it is the combination of the two that enables research that otherwise may be impossible to achieve (Hure et al., 2015).

2.7. Challenges of Administrative Data Linkage

Significant challenges to accessing and using administrative data for research purposes through data linkage are reported in the literature. These challenges were comprised of a range of issues across legal, ethical, technical and cultural concerns. The identified challenges, as presented in Table 2.8, will be explored under the themes and subthemes developed during the thematic analysis detailed in Section 2.2.1.

Table 2.8: Challenges of administrative data linkage

Theme	Subtheme
Consent	Informed consent required to enable research
	Can introduce bias
	Can reduce sample size
	Can compound section bias from original study sampling
	Varying rates of consent to data linkage
	Conflicting results on what influences consent to linkage
	Difficulty obtaining consent for data linkage
	Issues with attempting to link without consent
Unique identifiers	Need for unique identifiers across datasets
	Issues with depending on unique identifiers for matching
	Alternatives to unique identifiers
Data ownership and the role of data custodians	Unwillingness to share administrative data
	Resource implications for data custodians
	Lengthy approval processes
Quality and structure of administrative data	Data not collected for sharing/reuse
	Data not designed for research purposes
	Only covers a proportion of the population
	Data quality issues
Privacy and trust	Priority issue for participants
	Need for balance between privacy and research
	Difficulties quantifying privacy risks
Technology limitations	Lack of digitalised administrative records
	Need for better matching technologies

2.7.1. Consent

Informed consent is a process in which a participant is provided with all the details of a proposed project, including any potential risks and benefits, and has all the necessary information to make a knowledgeable and autonomous decision on whether or not they wish to take part (OECD, 2016, Thornton and Hipskind, 2017).

2.7.1.1. Informed consent required to enable research

The ethical and legal standing on consent in research is that voluntary informed consent from study participants should be the default method to enable data processing and, particularly in the case of medical research, consent remains the primary means of legitimating research (Lessof, 2009, Knies et al., 2012, Sala et al., 2012, Audrey et al., 2016a, Laurie, 2016). The OECD stated that "the default position should be that personal data is not collected, processed or shared without informed consent" (OECD, 2016; pp.15). Furthermore, it cannot be assumed that patients support the sharing of health data for research purposes as studies of patient perceptions suggest that seeking consent is preferred before using data for purposes other than direct treatment (Stone et al., 2005, Audrey et al., 2016a). However, making data linkage contingent on consent can be problematic. For these reasons, it is unsurprising that consent was the most commonly identified issue within this literature review.

2.7.1.2. Varying rates of consent to data linkage

The data linkage consent rates in longitudinal studies varied from 20%-90% and were dependent on the type of information being sought, with higher rates reported for education data linkage than health data and the lowest consent rates reported for linkage with financial data (McKinney et al., 2005, Tate et al., 2006, Sala et al., 2012, Sakshaug et al., 2013, Al Baghal, 2016).

2.7.1.3. Consent can introduce bias and there are conflicting results on what factors influences consent

The likelihood of a participant providing consent for linkage is dependent on many factors and if those participants who consent systematically differ from those who refuse, then selection bias can be introduced to the analysis and threaten the validity of the results (Tate et al., 2006, Walley, 2006, Kho et al., 2009, Lessof, 2009, Carter et al., 2010, ADT, 2012, Knies et al., 2012, Sakshaug et al., 2013, Sala et al., 2014, Salman et al., 2014, Al Baghal, 2016, Mostafa, 2016). Evidence from the literature demonstrated propensity to consent to data linkage is influenced by many participant factors including socioeconomic status, age, gender, and participants' perception of risk, privacy and altruism, as well as the household composition, the study environment and the interviewer (Sala et al., 2012, Al Baghal, 2016). However, the effects of these characteristics were not replicated in all studies with many reporting varying or contradicting results (Kho et al., 2009, Carter et al., 2012, Knies and Burton, 2014, Mostafa, 2016). From reviewing the literature, this potential for introducing bias which can undermine the research, appeared to be one of the major concerns with making data linkage contingent on consent.

2.7.1.4. Consent can reduce sample size and compound section bias from original study sampling

If the consent to linkage is low, meaning the complete study sample cannot be included in the linked research, then any results drawn from the analysis will no longer be representative of the wider population for which the sample is meant to represent (Brett and Deary, 2014, Sala et al., 2014, Mars et al., 2016, Mostafa, 2016). Also, if the characteristics influencing consent to linkage are similar to those influencing participation in the overall survey, any bias introduced in the original sampling frame can be multiplied in the linked data, further widening the gap between the research sample and the population it represents, meaning estimates are further biased by non-response and are therefore less generalizable (Tate et al., 2006, Kreuter et al., 2010, Sala et al., 2014, Al Baghal, 2016).

2.7.1.5. Difficulty obtaining consent for data linkage

Depending on consent may also limit research where the focus is on a specific population or condition which affects the ability to give informed consent. For instance, longitudinal studies which focus on ageing or older populations would expect a proportion of the participants to develop conditions such as dementia which would preclude them from being able to provide informed consent to linkage with their administrative records (Lessof, 2009, Brett and Deary, 2014). Similar challenges were experience with conducting research on minority populations with studies reporting lower consent rates to data linkage from these participants, meaning they are underrepresented in linked data research and results of analysis cannot be generalised to these populations (McKinney et al., 2005, Tate et al., 2006, Al Baghal, 2016). This would severely impede research on these population and important conditions such as dementia using linked datasets.

Also, longitudinal studies which begin in childhood, where initial consent was obtained from parents/guardians, will be required to acquire new consent for data linkage once the participants reach an appropriate age. If the children choose not to continue with the linkage that their parents had initially consent to, then this would affect the longitudinal analysis across the life course (Lessof, 2009, Brett and Deary, 2014). This uncertainty in accounting for consent across age transitions was highlighted as a major concern for researchers in a recent review of current longitudinal study practices in the UK (Townsley, 2016).

This issue can also be experienced with ongoing consent to data linkage. If a participant who has previously provided linkage consent is no longer in a position to confirm continuing consent or alternatively if a participant withdraws from a survey, consideration needs to be given to whether this consent can be used to justify ongoing linkage (Lessof, 2009). For example, if a participant had consented to annual linkage with the Hospital Inpatient Enquiry system (HIPE) should these annual linkages continue after a participant has withdrawn from primary data collection through interviews?

Another challenge with obtaining informed consent is that the nature of longitudinal studies means data is collected for long-term analysis, and often at the time of collection the study co-ordinators are unaware of potential future uses that may arise due new research theories or technologies. Therefore, truly informed consent at baseline collection, in which a participant is completely aware of the nature of data to be shared and all its potential future uses, is difficult if not impossible to achieve (Stone et al., 2005, Kaye et al., 2015). The OECD (2016) recommended that efforts should be made to obtain updated consent for these new and unforeseen uses as they arise. This can prove costly and time-consuming or even unfeasible particularly for historical samples where contact details for participants may be unavailable or outdated (Walley, 2006, ADT, 2012, Brett and Deary, 2014, Kaye et al., 2015). However, one potential solution suggested by the OECD is to focus consent on how data will be systematically processed and handled rather than on the specifics of original data collection (OECD, 2016).

2.7.1.6. Issues with attempting to link without consent

However, the role of consent in data linkage has not always been as salient and there are examples from historical studies where data linkage has occurred without consent. For example, a linkage study in the US matched nearly 1.5 million birth and foetal death records from 1980 to 1992 to establish individual timelines of women's reproductive histories (Adams et al., 1997). Additionally, obtaining consent is not always feasible, particularly when using data from a study that is no longer actively collecting data directly so has no contact with participants. Therefore, as argued by Lessof (2009) consent cannot be depended on as the unquestionable cornerstone of all potential research. In these situations, the OECD has recommended using ethics committees to review the risks and benefits and determine if the public benefits of any linkage would outweigh any potential risks to the individuals (OECD, 2016). This is a position supported throughout the literature and several studies identified from this review have overcome the issue of consent though ethical waivers (Stone et al., 2005, Jones and Elias, 2006, Walley, 2006, ADT, 2012, Brett and Deary, 2014, Mountain et al., 2016).

For example, Brett and Deary (2014) were attempting to carry out a data linkage project using the Scottish Mental Survey which consists of 75,252 participants born in 1936. The original study had not collected data or contacted the participants since the mid-1960s (Maxwell, 1969). The authors felt that attempting to seek consent for their follow-up linkage study would not be feasible due to the time that had passed and so instead applied to both the Privacy Advisory Committee and the Ethics and Confidentiality Committee of the National Information Governance Board to obtain a wavier on consent. As a result, the authors were able to link survey data collected from the participants between birth and age 27 with anonymised health records and the death registry.

A similar approach was used when linking survey data from the RAINE study with administrative data from WADLS (Mountain et al., 2016). Consent for the linkage was waivered through the University of Western Australia Human Research Ethics Committee for participants who were uncontactable or who failed to respond. If the analysis had been limited to only the participants who returned a consent form the response rate (RR) would have been 39%. However, as a result of the wavier, 1,697 additional participants were added to the linked analysis and only the 5 participants who returned a form specifically stating they did not consent were excluded from the project giving a RR of 98%.

However, despite the benefits that consent waivers brought for the above examples, the authors of both highlighted the long and complicated procedures required and cautioned against viewing the waiving of consent as an easy fix. To obtain permission for the linkage in the absence of consent, Brett and Deary (2014) were required to submit 210 separate documents to seven different regulatory bodies and the process took over 18 months to complete. The authors of the RAINE study similarly described the process as “laborious and lengthy” and took over a year to obtain final approval (Mountain et al., 2016).

These examples, combined with legal and ethical standards, suggests that seeking consent from participants should be still be considered first, and only bypassed when there is strong evidence to suggest depending on consent would significantly affect the analysis. Even in situations where consent is not sought, there must be efforts to protect the privacy of participants and ensure that the benefits of the linkage outweigh any potential risks for them individually.

2.7.2. Unique identifiers

A unique identifier is defined as code or tag which is associated with a single specified individual (Sariyar and Schlunder, 2016). One potential challenge to undertaking linkage projects in Ireland is the lack of a unique identifier (UI) used across all government services, particularly in the case of healthcare services.

2.7.2.1. The need for unique identifies across datasets

The presence of a UI, such as the National Health Service (NHS) number in the United Kingdom, is often used to enable linkage across datasets and facilitate effective matching (Hockley et al., 2008). In the absence of a UI to assist the matching process, attempting data linkage can be an onerous and resource intense task with high rates of potential error (Jones and Elias, 2006, Fredman et al., 2001). Many studies viewed the foresight of early governments to establish and include unique identifiers across administrative services as essential for the linkage research they have conducted (Holman et al., 2008, Hagger-Johnson, 2015). Additionally, using a UI means sensitive information such as names, addresses and DOBs do not need to be shared between data owners in order to facilitate matching and the presence of the UI across datasets enables deterministic or exact match linkage (Gill, 2001).

The ideal identifier has been described as “unique, universally available, fixed, easily recorded and at the same time readily accessible and verifiable” (Gill, 2001; pp.54). While not all government-issued UIs encompass every one of these characteristics, there are some which come close and are ideal for enabling matching across datasets, such as the NHS number in the UK which is unique, practically universal to the entire UK population and verifiable.

In a review of current longitudinal studies in the UK in order to develop a strategic plan for future research, Martin et al. (2006) highlighted that incorporating UIs such as National Insurance Numbers into longitudinal surveys would hugely enhance their research potential due to the vast amount of administrative data linkage that could be enabled through them. The benefits of using a high quality government issued identifier is they have very high discriminatory power as they are unique to each individual, the vast majority of the population will have been assigned one from interacting with government services and once assigned they tend not to change (Jenkins et al., 2008).

2.7.2.2. Issues with depending on unique identifiers for matching and potential alternatives

However, this requirement for a unique identifier was not consistent across studies and many had achieved successful matching in the absence of such an identifier. For example, the Office of National Statistics in the UK used surname, forename, postcode and any two parts of DOB to perform linkage between census data and other administrative datasets (Jones and Elias, 2006). In fact, Jenkins et al. (2008) argued that better matching could be achieved using other information and showed that linking longitudinal surveys and administrative datasets was more successful using a combination of sex, DOB, plus either postcode or first name and family name, yielding a raw linkage rate higher than depending solely on the participants' national insurance number. Some of the issues identified with UIs were participants consenting to linkage but not being able to provide their number, providing the wrong number and interviewers making transcribing errors when recording the number, all of which lower the number of possible matches with the administrative dataset. These results suggest that the lack of a UI across administrative datasets in Ireland should not preclude linkage projects from being attempted with existing surveys.

However, the authors also emphasised the weaknesses of depending of this data for matching as, even if a combination of variables are used, there may be low discriminatory power and a sufficient level of uniqueness may not be reached. For example, attempting to link a female survey participant named Mary Murphy would likely return several potential matches from a national administrative dataset. The authors also highlighted the importance of pre-processing when attempting linkage on these variables as there are often differences in recording conventions due to surveys' tendency to use participants' nicknames or truncated versions of official names which are usually used in administrative data. There is an added resource implication for this pre-processing but, in combination with matching algorithms, it can potentially improve the accuracy of the matching when UIs are not available (Jones and Elias, 2006, Jenkins et al., 2008).

While Jenkins et al. (2008) demonstrated that the best independent matching came from using a combination of sex, DOB and postcode, which had the highest raw linkage rate while also having both the lowest false-positive rate (matching records that are not actually the same person) and the lowest false-negative rate (indicating no match is present when there actually is a match), the authors concluded that the most effective matching protocol was to use a hierarchical technique using both UIs and the sex, DOB, postcode combination. This method was also endorsed by Calderwood and Lessof (2009), as depending solely on a UI can lead to false positives and negatives as the collection of these UIs during surveys is also subject to non-response and measurement error.

2.7.3. Data ownership & the role of data custodians

One challenge with data linkage which was repeatedly reported in the literature was a reluctance of data custodians to make data available for external parties, with the OECD describing this tendency not to share data as one of the major challenges in today's research environment (Holman et al., 2008, ADT, 2012, Brett and Deary, 2014, OECD, 2016).

A data custodian is defined as individuals, organisations, agencies or their representatives who are responsible for the collection and use of datasets, and have access to identifying demographic information such as name, address, DOB, etc. Most importantly, data custodians are responsible for ensuring the privacy of individuals is protected, in keeping with legislation, ethical guidelines and public interests (Kelman et al., 2002, Mountain et al., 2016).

2.7.3.1. Unwillingness to share administrative data

As emphasised by Robin (1992; pp. 1), there is widespread "underutilisation, misuse and non-use" of administrative data for secondary purposes. In many cases, this reluctance to share data for research purposes is born from fear that data which identifies individuals will become publically available and that the resulting implications will impact negatively on the data custodians. Also, the methods used during original collection may preclude sharing, such as statements of confidentiality in consent forms (Jones and Elias, 2006, Macleod et al., 2010, ADT, 2012, Brett and Deary, 2014, Moran, 2016).

An example of this reluctance to share data, in this instance with a wider linkage system rather than an individual study, was seen in the establishment of WADLS. The lack of key data owned by the national Australian Government rather than the local Western Australian authorities was described as one of the linking systems "most conspicuous shortcomings".

Brett and Deary (2014) cautioned that due to these difficulties with organisational and collaborative requirements, successful linkage systems are rare and most tend to operate at a local or regional rather than national level. In order to facilitate a national data linkage programme in Ireland, such as the model suggested by the HRB, it is clear that substantial inter-agency collaboration will be required (Moran, 2016). Two key successes highlighted by Holman et al. (2008) from the experience of WADLS was that the movement of senior staff between academia and government departments helped to establish common goals and foster negotiations. Also, the housing of the data linkage unit, which comprised academic staff from WADLS, within the government departments overcame privacy concerns by enabling the linkage process without the need for identifiable administrative data to leave government offices.

2.7.3.2. Resource implications for data custodians

In other cases, data custodians felt the resources which they had invested into the original collection justified retaining ongoing control over the data (Kelman et al., 2002, ADT, 2012). Personal health data in particular is viewed as a valuable commodity, which influences custodians' willingness to share the data (HIQA, 2010). There are also considerable resource requirements for the custodians in making their data available for secondary use which can have implications on their willingness to share data (Jones and Elias, 2006).

2.7.3.3. Lengthy approval processes

In one case study reviewed by the Administrative Data Taskforce (2012) in the UK, a longitudinal survey had obtained written permission from the participants to access a specific subset of their administrative data but the process for the custodian to make this data available took more than two years which significantly delayed the analysis of the data.

Returning to the WADSL example from Section 2.7.3.1, negotiations to incorporate the national Australian Government data began in 1998 but pilot linkage was not attempted until 2001 and linked data was not available for researcher access until 2005. Even when the data became available, the original custodians retained ownership and continued to control access and use of any datasets that contained their original data (Holman et al., 2008).

Even where data is available for research purposes there is usually a lengthy approval process to obtain access and often a requirement for researcher to obtain separate approval from each administrative or government department with ownership over the data (ADT, 2012, Brett and Deary, 2014). This is the case for WADLS, with researchers required to apply for separate approval from each data custodian (Holman et al., 2008). Similarly there are often additional restrictions or tighter access controls for particularly sensitive data (Gissler and Haukka, 2004).

2.7.4. Quality and structure of administrative data

A challenge identified from the literature is that administrative data is not collected in a structure designed for research and the quality of the data cannot be assumed. Martin et al. (2006) strongly advised against viewing administrative data as a panacea for the weaknesses of longitudinal survey data.

2.7.4.1. Data not collected for sharing/reuse

Jones and Elias (2006) stressed that one barrier to fully exploiting data linkage is the limited and fragmented knowledge of administrative data resources among the academic and wider research community. In Ireland, the administrative data infrastructure developed in an uncoordinated manner which has led to a fractured information structure (HIQA, 2017c). There is inadequate information available on the range of administrative data being collected and maintained or about the scope and potential of these resources. Even where administrative data is made available, researchers reported difficulty using the data due to unclear structures and a lack of accompanying metadata to make data fully understandable for secondary users, which can result in underutilisation or misuse of the provided data (Jones and Elias, 2006).

2.7.4.2. Data not designed for research purposes

One of the main issues encountered is that administrative data is collected primarily for service-delivery and therefore the data is not always structured for secondary use. As this data is not generated for research purposes, it is generally not in a format that can be directly incorporated into existing survey data and can require significant cleaning and preparation in advance of any use (Jones and Elias, 2006, Calderwood and Lessof, 2009, Brownell and Jutte, 2013, McGhee et al., 2015, Townsley, 2016). Some of the issues reported in the literature were varying collection processes across departments, lack of validated coding standards or terminology systems, transcribing errors and mis-recording (Jones and Elias, 2006).

Additionally, as administrative data is not collected with any statistical motive, it is not possible to specify or modify the collection methods to meet statistical standards or research needs. Researchers cannot add or change the variables that are collected and cannot insist on strict study protocols being implemented during collection which would be standard during equivalent study collection (Jones and Elias, 2006, Brownell and Jutte, 2013, Sakshaug et al., 2013, McGhee et al., 2015).

2.7.4.3. Only covers a proportion of the population

The administrative data will also be limited to those who interact with the administrative services (Fredman et al., 2001, Brownell and Jutte, 2013). For instance, healthcare records will only be available for those who consult healthcare services (Knies et al., 2012, Brett and Deary, 2014). Therefore, there is a risk that undiagnosed conditions or unmet healthcare needs will not be reflected in the administrative records.

2.7.4.4. Administrative data quality issues

It is also vital that the quality of the administrative data is reviewed and carefully appraised prior to any data linkage, particularly if the purpose of the linkage is to address the survey data quality issues identified in Section 2.6.1 (Martin et al., 2006). Administrative data is subject to similar data quality issues such as missing and inaccurate values and is heavily dependent on the nature and means of collection. Historical data such as birth records which were collected manually, for example, may not be subject to the same collection and validation standards as modern electronic care records and the data would require cleaning, coding and digitising before linkage could be attempted (Jones and Elias, 2006, Soloff et al., 2007, Calderwood and Lessof, 2009, Brett and Deary, 2014, Hure et al., 2015, Carroll et al., 2016).

2.7.5. Privacy and trust

There is an inherent expectation that data will be kept confidential when accessing administrative services such as healthcare (Brett and Deary, 2014). Privacy is an essential right due to its "central role in enabling people to define, develop, and maintain their personal and social identities" (OECD, 2016; pp.8).

2.7.5.1. Need for balance between privacy and research

Particularly in the case of healthcare data, respect for privacy and confidentiality is essential for maintaining the patient-healthcare provider relationship (Stone et al., 2005) and this expectation of privacy is supported by legislation and ethical guidelines. Additionally, data custodians are also subject to industry specific standards, such as the Caldicott Principles, which are designed to safeguard patient-identifiable data in the UK (Caldicott, 2013).

As result, one of the frequently cited barriers to data linkage related to potential invasion of privacy and the need for data matching processes which effectively protect the anonymity of participants, predominantly in the case of digital data (Townsley, 2016). In many cases, authors reported that a fear of the perceived risks to privacy was limiting willingness to share data and that the potential benefits of data linkage were not being exploited, resulting in delays and wastage for research projects, particularly in the health domain (Flowers and Ferguson, 2010, Laurie and Sethi, 2011, Brett and Deary, 2014, Salman et al., 2014). Data custodians were primarily concerned with disclosure risks, "the releasing of confidential information relating to individuals (or businesses) which breaches the legal obligations of confidentiality" (Jones and Elias, 2006; pp.75) and it was reported that preventing this disclosure often conflicted with the advancement of research through data sharing (Stone et al., 2005). However, there is justification for this cautiousness in relation to privacy, as any intrusion on privacy by one linkage project will negatively impact trust in any future projects and will serve to undermine any long-term linkage strategies (Al Baghal, 2016, OECD, 2016).

Conversely, Holman et al. (2008) argues that using data linkage positively impacts participants' privacy through best practice protocols and safeguards which address privacy concerns and allow data custodians to retain control over their data. This is achieved by eliminating the need to share data together with personal identifying details, instead the data is split during the matching process and only the data custodian retains the linkage key. This method of privacy conservation is employed by ELSA in order to link to NHS primary healthcare records (Appendix A). Separating the data of interest from identifiers in this way greatly reduces the disclosure risk and results in increased privacy protection when sharing data (Martin et al., 2006, Mountain et al., 2016). The use of persistent unique identifiers across administrative datasets further reduces the disclosure risk as it negates the need to share common identifiers such as names and addresses (Holman et al., 2008).

However, as highlighted by Jones and Elias (2006), the lack of direct identifiers such as names does not means data is anonymous as, due to the nature of the data contained within administrative records, there is still a significant risk of deductive disclosure and it is often possible to identify participants by combining information from several variables. In a qualitative study of participants' opinions on data linkage, Audrey et al*.* (2016a) showed participants were not confident about the effectiveness of anonymisation and did not believe it negated the need for consent for data linkage.

2.7.5.2. Priority issue for participants

Maintaining trust is particularly important in longitudinal studies due to the long-term commitment required from participants. If there is a perception that privacy and confidentiality are not being respected by researchers, there is likely to be a negative impact on sample retention. It is therefore essential that potential linkages are reviewed in advance of requesting linkage consent from participants, to ensure the research benefits outweigh any perceived intrusion for participants (Martin et al., 2006). Studies by Sakshaug (2012) and Korbmacher and Schroeder (2013) have shown that participants who have concerns about the privacy of administrative records are less likely to consent to data linkage within a survey. Furthermore, Sala et al*.* (2012) demonstrated that a participant's attitude to privacy has a stronger effect on propensity to consent than either their demographic or socioeconomic characteristics.

2.7.5.3. Difficulties quantifying privacy risks

A noticeable difficulty is that it is not possible to effectively quantify disclosure risk or its impact on privacy as it is dependent on many varying factors, such as the size and structure of the dataset, the nature of the data and the types of variables, making it challenging to balance the potential risks against the benefits of data linkage. This lack of clarity has caused data custodians to be overly cautious and delayed or prevented data being made available for linkage with research data (Jones and Elias, 2006).

In order for the benefits of data linkage to be realised, issues surrounding privacy and confidentiality need to be addressed to lessen the reluctance of data custodians making data available for linkage and to ensure confidence and trust in any potential research projects (Martin et al., 2006).

2.7.6. Technology limitations

The increased use of administrative data for research purposes has been greatly facilitated by advances in technology and the computerisation of administrative records (Holman et al., 2008, Calderwood and Lessof, 2009). However, technological limitations remain an ongoing barrier to fully exploiting data linkage in Ireland.

2.7.6.1. Lack of digitalised administrative records

One significant issue preventing further data linkage, is that many administrative data sources are still maintained in local paper records without any central systems in place. This will result in a digital bias in the data linkage that is occurring and some administrative data will be excluded from linkage projects completely, due to the additional work that would be involved in digitising them or will result in considerable costs to individual studies that pioneer their use (Hagger-Johnson, 2015, Robin, 1992).

2.7.6.2. Need for better matching technologies

Jenkins et al. (2008) highlighted a significant need for improved matching technologies, such as the development of pre-processing software to assist with data cleaning in advance of matching, as well as advanced probabilistic matching algorithms. It was strongly recommended that investment in resources such as these is conducted centrally to ensure consistent protocols and to prevent individual studies baring disproportional costs which may stall advances in the area (Fredman et al., 2001, Jones and Elias, 2006, Holman et al., 2008, Jenkins et al., 2008, Boyd et al., 2014, Townsley, 2016).

2.8. Key Findings from the Literature

This review found that there are several benefits of administrative data linkage with practical examples demonstrating that realising these benefits is achievable. Using administrative data to enrich survey data was described as one of the primary motivations for utilising data linkage in order to increase research potential and also to enhance the value of existing longitudinal survey data. Data linkage also had benefits in relation to addressing key longitudinal methodological issues such as sample tracking, attrition and participant burden, and measurement error. Additionally, data linkage can enable cost savings and efficient use of resources for both the research community and the administrative data custodians.

Despite the potential benefits of incorporating administrative data linkage into research studies, the literature review uncovered several barriers to its implementation. Legislative issues in relation to data protection and consent can significantly impact any use and analysis of the linked data due to the introduction of bias. Privacy concerns and a reluctance of data custodians to engage with researchers can prevent data being shared and, even when data is made available, technical and practical challenges come to the fore, such as a lack of UIs and appropriate linking technology, and data quality issues, which can impact its potential use.

One of the key points identified from the review is that administrative and survey data have differing strengths and weaknesses and it is the combination of the two data types which has the potential to significantly contribute to the research environment. For example, while administrative data may not be subject to the same level of recall and attrition bias, surveys can achieve finely detailed data which would not be possible through administrative services such as health perceptions and unmet healthcare needs of participants, in addition to collecting this data within strict study protocols.

2.9. Conclusion

This chapter presented the results of a comprehensive review of the available international literature in relation to data linkage with a focus on the benefits and challenges of performing linkage with longitudinal research studies. The available literature demonstrates that there are many benefits to data linkage including data correction and enrichment as well as controlling participant burden and attrition and reducing costs. While this data linkage does raise many challenges such as obtaining appropriate consent, achieving accurate matching and custodians' reluctance to engage, the evidence from the literature establishes that it is possible to achieve effective linkage projects between longitudinal studies and administrative datasets.

The information obtained from this literature review is used to compare against and evaluate the current experiences of longitudinal researchers in Ireland. The identified benefits and challenges will be explored in more detail in Chapters 6 and 7.

3. Chapter 3: Relevant Legislation, Standards and Ethics

3.1. Introduction

As highlighted by Chassang (2017), personal data is critical for enabling high quality and reliable scientific research. However, there are legislative and regulatory guidelines which must be complied with in order to ensure the privacy and other rights of participants are protected and, while there has been an increase in the number of research projects utilising administrative data, there has been a corresponding increase in the legislative, regulatory and ethical controls surrounding the process (Calderwood and Lessof, 2009). This has resulted in a complex setting of unclear and conflicting requirements (Laurie and Sethi, 2011). For example, Brett and Deary (2014), cautioned that a lack of understanding of the legal and ethical framework within which data linkage must occur significantly delayed their project.

The main legislation which relates to data linkage efforts revolves around data protection, but industry standards and ethical guidelines must also be considered. In order to assess how these would impact any future administrative data linkage and to inform the development and organisation of a national data linkage infrastructure, the relevant legislation and regulations are reviewed and summarised, with a particular focus on impending changes.

3.2. Data Protection Legislation

3.2.1. Current data protection legislation

Privacy has long enjoyed a protected stature in Irish law as the right to privacy is enshrined in the Irish Constitution by Article 40.3.1°, which implies a right to privacy through the protection of personal rights (Kelly, 2003). Currently, the key legislative controls are established through the Data Protection Act, 1988 and the Data Protection (Amendment) Act, 2003. These current Irish laws are borne from the European Data Protection Directive and focuses around eight key principles (Table 3.1) that allow data controllers to identify and implement the necessary measures to remain data protection compliant (DPC, 2017).

The current legislation allows for certain exemptions when processing data for research purposes. For instance, it is legally compliant to retain data for longer than is necessary for the specified purpose if it is kept for research purposes (Lambert, 2016, OECD, 2016). Additionally, the Data Protection Commissioner (DPC) has published specific guidelines on the use of personal data for research purposes (DPC, 2007).

Table 3.1: Eight principles of data protection

1. Obtain and process information fairly
2. Keep it only for one or more specified, explicit and lawful purposes
3. Use and disclose it only in ways compatible with these purposes
4. Keep it safe and secure
5. Keep it accurate, complete and up-to-date
6. Ensure that it is adequate, relevant and not excessive
7. Retain it for no longer than is necessary for the purpose or purposes
8. Give a copy of his/her personal data to an individual, on request

Source: Data Protection Commissioner (2017)

There are also specific pieces of legislation such as the Infectious Disease Regulation, 1981 and the National Cancer Registry Board (Establishment) Order, 1991 and their subsequent amendments, which allow for the collection and processing of specific personal data in Ireland. For example, consent from patients is not required for their personal details and details of their cancer diagnosis and treatment to be added to the national register as the National Cancer Registry Board (Establishment) Order, 1991 allows relevant data be collected directly from their patient records. These specific acts are often used to give legal protection to data processing which is necessary to protect public health. For example, the Infectious Disease Regulation, 1981 allows the Health Services Executive (HSE) to monitor disease incidence meaning emerging outbreaks can be identified.

3.2.2. The General Data Protection Regulation

However, the European Data Protection Directive is due to be replaced in May 2018 by an EU-wide General Data Protection Regulation (GDPR)[1]. Regulations and directives from the EU differ as, while a directive is a binding instruction which must be enacted through legislation in each

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] Official Journal L 119/1

Member State with the clauses of the directive being incorporated and adapted to suit local legislation, an EU regulation applies directly within all EU Member States (Schutze, 2015). The introduction of the GDPR strongly reaffirms the EU's focus on protecting individuals' privacy and will have implications for the Irish research community, requiring researchers to adapt their practices to ensure they are compliant with the new legislation.

The aims of the eight principles of the current Directive (Table 3.1) are maintained in the GDPR but they are incorporated into processing concepts and new general principles have been added. While, the focus underlying these new principles were included in the Directive, they have been strengthened, clarified and highlighted in the GDPR (Chassang, 2017).

One of the new principles that will need to be incorporated into the work of researcher is that of accountability. There will be an onus on data controllers, including researchers that collect and use personal data, to be able to demonstrate compliance with the GDPR. Researchers will be required to maintain written records of any data processing that they perform which falls under the remit of the GDPR, and be able to produce these as evidence of compliance to the Office of the Data Protection Commissioner (Article 30).

There is also a new principle which focuses on data protection by design which highlights the need to incorporate data protection into the design of every project and ensure it is core during the complete lifecycle of data processing (Article 25). This principle indicates the important combined role of technology and organisational measures to ensure compliance: the available technology, the cost of implementation, the nature and scope of the processing and the potential risks to the data subjects should all be reviewed in advance in order to effectively establish privacy by design for a project. For the research community, compliance with this data protection by design principle may be less onerous than for other sectors as justification for data collection and specifying and defending data protection procedures are already required for most grant and ethics applications in advance of beginning any project (Chassang, 2017).

3.2.3. Processing for research purposes within the GDPR

The GDPR does, however, provide clarity for the research community with processing for research purposes forming the basis for six Articles and fourteen Recitals, across definitions, exemptions and safeguards. Notably, the GDPR includes Article 89 which is dedicated to issues relating to the processing of data for research purposes and clarifies how research can qualify for exemption or derogation from the principles of the Regulation and what safeguards must be applied. The GDPR has also established a legal definition of research and, in order for data linkage with longitudinal studies to benefit from the exemptions allowed for research purposes, the linkage processes must aligned with these definitions (Box 1).

Box 1: Definitions established in the GDPR

Scientific research (Recital 159)**:** For the purposes of this Regulation, the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research… Scientific research purposes should also include studies conducted in the public interest in the area of public health.

Statistical purposes (Recital 162)**:** mean any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.

The GDPR incorporates several exemptions for data processing for research or scientific purposes which have the potential to impact administrative data linkage in Ireland:

- Under the GDPR, the processing of sensitive personal data, such as healthcare information, is prohibited (Article 9.1). However, processing for research is included as one of the exemptions of this rule through Article 9.2(j), meaning, while consent of data subjects is still the preferred justification for processing sensitive personal data, it is possible to proceed without consent, if the processing is for research purposes, provided appropriate safeguards are implemented to ensure the fundamental rights of the data subjects are protected. The safeguards adopted by researchers must ensure that technical and organisational procedures, such as data minimisation and pseudonymisation, are sufficient and adequate.

- As stated in the GDPR, informed consent requires a data subject be made aware of all intended data processing at the time of consent (Article 7). One important concession acknowledged in Recital 33 of the GDPR is that when collecting data for scientific research purposes it is often not possible at the time of collection to know exactly how the data will be used and so currently studies often depend on broad consent from participants (Chassang, 2017). This is particularly relevant for data linkage, as once survey and administrative data have been linked it may be used to address several different research questions or initial planned analysis may lead to future research hypotheses. Recital 33, therefore, states that research participants should instead be allowed to consent to areas of research which are compliant with recognised ethical standards, meaning any use of the data will be limited to ethically approved research. This an important development as it gives legal standing to respecting ethical standards of the research community (Chassang, 2017). Additionally, participants should be allowed to consent separately to individual sections of a research project insofar as possible; for example, participants should be asked for separate consent to link to each administrative data source rather than a blanket consent for linkage to any data.

- The GDPR also provides clarification around data erasure in research studies, providing a legal basis for retaining research data when a participant has withdrawn consent, if deleting the data is likely to seriously impair the research (Recital 65; Article 17.3(d)).

- As part of Recitals 52 and 53, research purposes are one of the included justifications for establishing Member State laws to enable the processing of sensitive personal data particularly health data. This would allow Ireland to introduce specific legalisation that would provide a legal standing for a data linkage system such as the DASSL model, recommended by the HRB (Section 3.4), and allow for the inclusion of health data in in any potential linkage systems (Moran, 2016).

- Recital 157 is particularly relevant for administrative data linkage as it provides clarification on ‘coupling information’ from registries for social science purposes and its ability to provide high-quality information which has the potential to improve the efficiency of services and enhance quality of life for individuals (Box 2). The Recital states that personal data can be processed in this manner to facilitate enhanced research, provided the processing is subject to appropriate conditions and safeguards established through Member State law.

Box 2: Recital 157 of the GDPR

By coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer and depression. On the basis of registries, research results can be enhanced, as they draw on a larger population. Within social science, research on the basis of registries enables researchers to obtain essential knowledge about the long-term correlation of a number of social conditions such as unemployment and education with other life conditions. Research results obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people and improve the efficiency of social services. In order to facilitate scientific research, personal data can be processed for scientific research purposes, subject to appropriate conditions and safeguards set out in Union or Member State law.

3.2.4. New procedures established in the GDPR relevant for research

Rather being prescriptive, the GDPR establishes a risk-based approach to data protection which can be adapted for and applied to the various types of and environments in which data is processed. To facilitate this adaptive approach the GDPR introduces some new procedures which can be applied to the research community (Chassang, 2017).

One of the main concepts introduced by the GDPR which may affect research practices is the necessity for a dedicated Data Protection Officer (DPO) in organisations that process data (Articles 37, 38, 39). A DPO should exemplify expert knowledge in both data protection legislation and practice and should independently advise on and monitor internal compliance with the GDPR and act as a contact point for data subjects and the DPC. Article 37 and Recital 97 give clarity on the situations in which a DPO is necessitated and it would appear that longitudinal studies fall within this remit based on the inclusion criteria of processing carried out by public authorities or bodies, processing that requires regular and systematic monitoring of data subjects on a large scale and processing of special categories of data, such as health data, on a large scale. Any longitudinal study attempting to embark on data linkage would be required to consult with their DPO in advance of any processing and ensure that recommendations made by the DPO are complied with for any processes relating to personal data.

Additionally, the GDPR establishes a legal requirement for completing Data Protection Impact Assessments (DPIA) which act as a self-assessment tool for identifying potential risks of data processing, determining their severity and developing measures to address them prior to any processing being undertaken (Chassang, 2017). Article 35 states that, where any processing is likely to result in significant risk for data subjects, a DPIA is required in advance to identify any potential risks of the proposed processing. The DPIA should include a systematic description of the proposed processing including the reason for the processing, an assessment of the necessity for the processing, and assessment of the potential risks to data subjects and the measures intended to mitigate these risks and maintain compliance with the GDPR such as technical and operational safeguards.

Notably, if a DPIA identifies any potential high risks for data subjects, regardless of whether or not the proposed measures would mitigate these risks, the DPC must be consulted prior to any processing occurring. The DPC will then review to ensure the processing is compliant with the GDPR and that the proposed control measures sufficiently mitigate the risk to the data subjects. This represents a key change in how data controllers such as researcher interact with the DPC as, rather the current system of registering all processing activities, they are now required to carry out initial self-assessment through the DPIA and obtain approval for processing which represents a high risk to data subjects (Chassang, 2017).

This new process will potentially impact researchers attempting to link with administrative data as they will now be required to carry out a DPIA and, if potential risks are identified, will need to notify the DPC. Additionally, if the risks of the linkage cannot be sufficiently mitigated, researchers will be prevented from continuing by the DPC. These changes may also impact funding applications, as there is potential for projects that have received funding being prevented from continuing by the DPC. Therefore, completion of DPIAs and obtaining DPC approval in advance of grant applications may become the established practice (Chassang, 2017).

3.3. HIQA Information Management Standards

Many of the potential administrative data sources which could be utilised for data linkage in Ireland fall under the remit of HIQA, and some of the longitudinal studies themselves can be classed as national data collectors due to the nature of the data they accumulate. Therefore, any potential linkage projects need to maintain compliance with HIQA's Information Management Standards (HIQA, 2017c). These recently launched standards complement the existing National Standards for Safer Better Healthcare which were launched by HIQA in 2012.

The new standards must be adhered to by all data custodians identified in the HIQA Catalogue of National Health and Social Care Data Collections (HIQA, 2014a). The new framework consists of ten standards which sit within six overall information management themes (Table 3.2). It is essential that any proposed linkages are compliant with these standards as HIQA will also be assessing and monitoring compliance. HIQA have also developed a self-assessment tool so data custodians can review their performance against the standards. This self-assessment tool will form the first step of HIQA's review programme (HIQA, 2017b).

These new standards will provide guidance and direction for researchers when processing health and social care data, particularly in relation to information governance which, evidence from the literature suggests, has been unclear, varied and often not documented (Stone et al., 2005, Brett and Deary, 2014, Hagger-Johnson, 2015). This requirement for documented practices is addressed in HIQA's prerequisite that projects have statements of both information practices and data quality as well as data quality frameworks which detail issues such as data policies and procedures, data performance indicators and quality metrics (HIQA, 2017a).

The standards also echo the requirement for PIA which is a requirement under the GDPR, meaning completing any linkage project without first completing a PIA with violate both legislative and information standard requirements.

The need for public engagement and transparency is also enshrined in the standards as all statements and documentation must be made publically available (HIQA, 2017c). Such public engagement was highlighted as a key step in establishing successful linkage projects in other countries (Holman et al., 2008, ADT, 2012, OECD, 2016), and its importance in the Irish setting for potential linkages is seen in its inclusion in the DASSL model (Moran, 2016). The data quality statement, which details the data's specific strengths and weaknesses including accuracy, completeness, reliability and validity, must be included in all published outputs from the project (HIQA, 2017a). This would impact any publication of data analysis based on linkage using data sources which fall within the remint of the HIAQ standards.

Table 3.2: HIQA Information Management Standards

Theme 1: Person-centred	
Standard 1:	The managing organisation of the national health and social care data collection has effective arrangements in place to protect the privacy of people about whom it holds information.
Theme 2: Governance, Leadership and Management	
Standard 2:	The managing organisation of the national health and social care data collection has effective governance, leadership and management arrangements in place, with clear lines of accountability to ensure that its objectives are met.
Standard 3:	The managing organisation maintains a publicly available statement of purpose that accurately describes the aims and objectives of the national health and social care data collection.
Standard 4:	The managing organisation of the national health and social care data collection is compliant with relevant legislation and codes of practice.
Theme 3: Use of Information	
Standard 5:	The managing organisation of the national health and social care data collection complies with health information standards and nationally agreed definitions to enable comparability and sharing of information
Standard 6:	The managing organisation of the national health and social care data collection systematically assesses, monitors and improves the quality of the data it holds to ensure its objectives are met.
Standard 7:	The managing organisation of the national health and social care data collection disseminates data and information appropriately and ensures that data users can access data and information in a timely manner to meet their needs.
Theme 4: Information Governance	
Standard 8:	The managing organisation of the national health and social care data collection has effective arrangements in place for information governance which ensure that personal information is handled legally and securely.
Theme 5: Workforce	
Standard 9:	The managing organisation of the national health and social care data collection plans, organises and develops its workforce to effectively deliver its objectives.
Theme 6: Use of Resources	
Standard 10:	The managing organisation of the national health and social care data collection plans and manages the allocation and use of resources to ensure its objectives are met.

Source: HIQA (2017c)

3.4. HRB Data Access, Storage, Sharing and Linkage (DASSL) Model

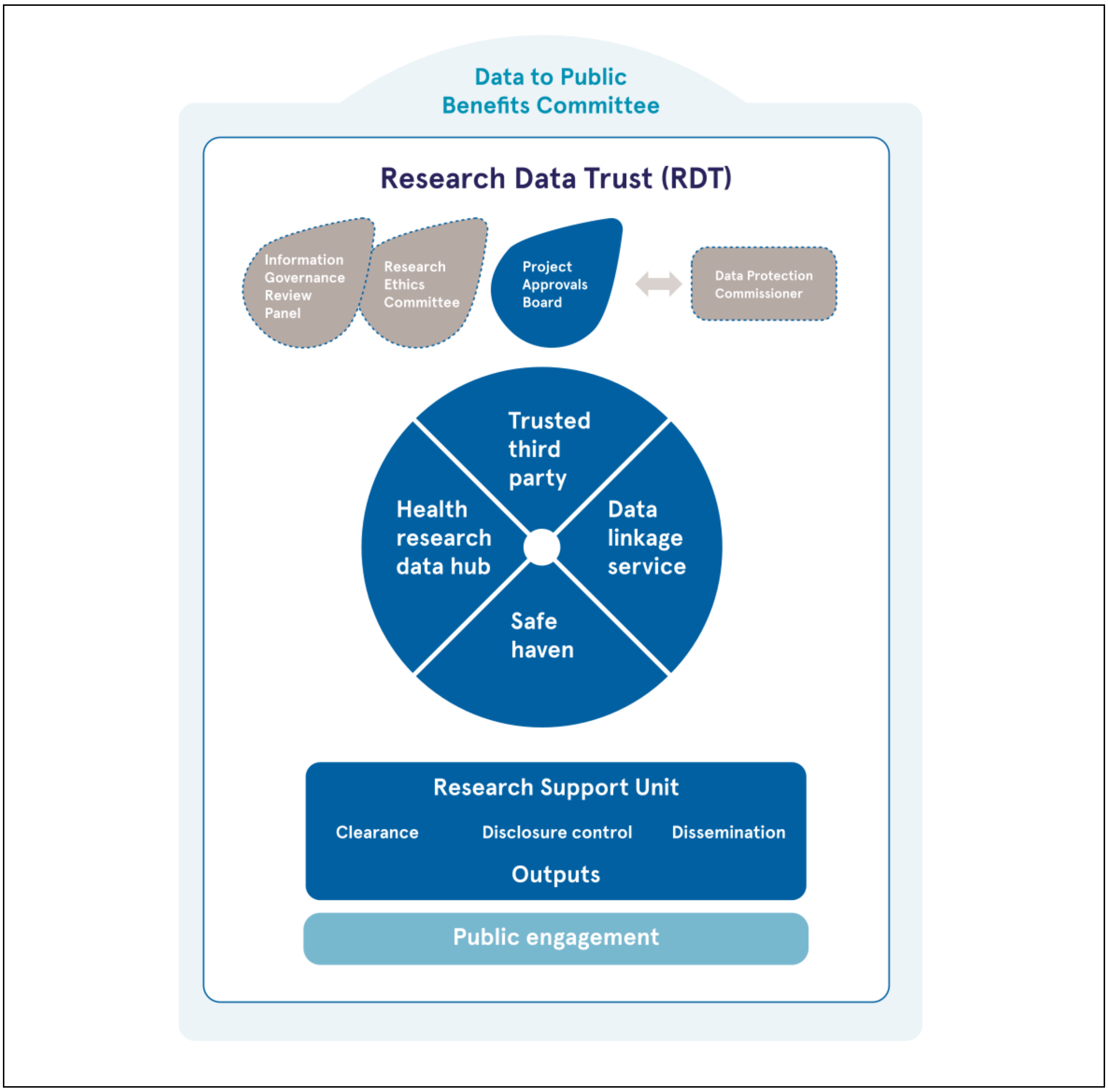
In an effort to promote a data environment which enables health-related research in Ireland, the HRB developed a proposed model to address data access, storage, sharing and linkage. The development of the model followed a robust review of international practices and stakeholder engagement and the final model aimed to address the concerns and needs of researchers by identifying the required services and infrastructure needed to enable safe and effective DASSL (Moran, 2016). The potential of the DASSL model to address the challenges of data linkage will be discussed further in Section 6.4.2.

The proposed model includes a governance structure, practical linkage facilities, research support and public engagement elements (Figure 3.1) and aims to provide the technical and intuitional environment required to enable "the 5 safes" of data access and use (Desai et al., 2016): safe projects (valid research purpose); safe people (trusted researchers); safe data (people's data protected); safe setting (security controls and secure environments); and safe outputs (effective disclosure control). The model evolved from seven key elements, detailed in Table 3.3, which would be overseen by a proposed Research Data Trust (RDT):

Table 3.3: Elements of DASSL model designed to facilitate research

DASSL Element	Details
Governance	Provides information governance to overall model and also to individual projects facilitated by DASSL. Incorporates ethics committee to approve submitted projects.
Health Research Data Hub	Facilitates safe access to data for approved projects in accordance with set governance procedures.
Third-party data linkage service	Securely links and integrates data from different custodians. Once linked, anonymised data is made available to approved researchers.
Save setting/haven data access	Secure 'locked-down' environment within which researchers can access linked datasets.
Research Support Unit	Provides guidance, training and assistance to researchers, ensuring data is used appropriately and supports custodians in providing well-documented data.
Output checking and disclosure control	Reviews all data output to ensure individuals cannot be identified.
Public engagement	Ensures and facilitates ongoing education, consultation and engagement with the public in relation to the use of their data.

Source: modified from Moran (2016)



Source: Moran (2016)

Figure 3.1: HRB DASSL model

3.5. Research Ethics

As well as the legislation and regulatory standards described above, researchers are also governed by ethical standards. While ethics and legislation can complement each other, they perform different roles in respect to research. Whereas the structures described in Sections 3.2 and 3.3 dictate what must and must not be done in order to maintain compliance, ethical principles focus on whether research should be undertaken and how it should be conducted in accordance with best practice (OECD, 2016).

Ethics are concerned with the integrity of research, defined as the "active adherence to the ethical principles and professional standards essential for the responsible practice of research" and address issues such as research misconduct, conflicts of interest, fabrication, falsification, plagiarism, confidentiality, data management, responsible data sharing and appropriate interaction with study participants (Chassang, 2017; pp.3, ADT, 2012, Council of the EU, 2015). Ethical concerns associated with secondary use of administrative data were identified by Audrey et al. (2016b) as an important issue for researchers attempting to utilise data linkage.

As detailed by Chassang (2017), ethical issues fall outside the realm of the data protection legislation, however, within the GDPR consideration is given to how ethical guidelines provide practical guidance to scientific research and Recital 33 states that when depending on research purposes for exemptions to consent, all processing should be in keeping with recognised ethical standards. Additionally, some research falls completely outside the remit of data protection legislation, such as that involving deceased participants or that using data which is anonymised, but it is vital that it is still conducted within ethical guidelines in order to protect participants and maintain the integrity of the wider research community (OECD, 2016).

Research ethics fall into two general categories; firstly, how researchers interact with each other and their own methodological standards through dissemination of research and peer reviewing, and secondly, how researcher interact with the surrounding world such as with study participants. Both categories of principles are necessary to achieve ethically sound research (OECD, 2016). In 2016, the OECD launched updated ethical principles in response to the increasing digitalisation and reuse of routinely collected administrative data for research purposes (Table 3.4).

The aim of these principles is to provide a framework for an ethical approach to research. Many of the principles echo the requirements of the incoming GDPR, such as the importance of consent (Principle 2), clear purpose for processing (Principle 3) and assigned responsibility roles (Principle 6). Similar to the HIQA standards, there is also a strong emphasis on data quality (Principle 4) as well as transparency and public engagement with the necessity to make publically available information on how data will be used, shared and protected (Principles 1 and 3). As can be seen from Principle 5, there is a requirement to review any potential research projects in advance to identify and address any negative impacts for participants. This is similar to the impact assessment requirements seen in both the GDPR and the HIQA standards.

Additionally, these ethical principles address how researchers interact with participants and how this may influence the balance of power between the two sides (Principle 7). The importance of ethical review bodies, which have the capacity to review and approve potential linkage projects with the necessary independence and expertise, is also included (Principle 8).

Table 3.4: OECD ethical principles

1. Mechanisms for the safe and responsible sharing of personal data, including mechanisms for the protection of privacy of data subjects as well as for public input and accountability, should be established and made public by data owners/controllers. Data should be shared as openly as is feasible within the relevant legal and ethical constraints.

2. The default position should be that personal data is not collected, processed or shared without informed consent. Efforts to update consent for new and unanticipated uses should be made where feasible.

3. Clear articulation of purpose should be provided before a research project using personal data is carried out. In many instances, this will entail the development of transparent long-term plans and mechanisms for communicating any updates.

4. With a view to both the impact of the research and respect for data subjects, data quality should be considered to ensure that it is fit to fulfil the stated research purpose.

5. Before a research or data collection project is undertaken, care should be taken to consider potential negative impacts, for individuals or groups, arising from the proposed project. Any potential negative consequences should be weighed against societal benefits, taking account of any mitigating actions to reduce the risk or impact of potential negative consequences.

6. Unambiguous distribution of responsibilities should be agreed in advance of any research-related data handling.

7. Data holders, research funders and researchers have a responsibility to consider how their role in a proposed research project would contribute to the balance of power and influence between their institutions and individual data subjects.

8. Data holders and research institutions should ensure they have access to an ethics review body (ERB) with the capability to review proposals to use data for research

Source: OECD (2016)

3.6. Conclusion

It is crucial that all longitudinal researchers are aware of the relevant legislation and regulations which will affect any potential data linkage projects. Understanding of and compliance with legislation, standards and ethics are all required to ensure responsible research processes and practices. For example, if researchers are not trained in relation to the incoming GDPR or cannot consult with an appropriate expert, they may conduct linkage without completing a DPIA in advance and such non-compliance would leave them subject to significant fines from the DPC.

The emerging issues detailed in this chapter highlight that researchers will need to become more accountable for their data management and governance and begin to build or enhance their data protection, information management and ethical capabilities. However, it is intended that these changes will likely have a positive impact on research practices including the planning and conducting of any potential data linkage projects.

4. Chapter 4: Research Methodology

4.1. Introduction

This chapter describes the methodology utilised to address the research question presented in Section 1.2, including the rationale for the chosen methods, the collection tools used and any ethical and data quality concerns.

The aim of the research detailed in this chapter is to identify existing examples of and demand for administrative data linkage by longitudinal researchers and to review the benefits and challenges experienced in order to further the understanding of the current situation in Ireland.

4.2. Research Question

The research question *what are the benefits and challenges of linking health and administrative data with research data in Ireland* will be addressed by reviewing the existing international literature and exploring the current research environment for linkage with longitudinal research data in Ireland.

This question was addressed using a multistage process incorporating:

1. A review of existing literature on the benefits and challenges to data linkage with a focus on those encountered in the longitudinal research environment.
2. Conducting a survey of researchers working on Irish longitudinal studies to identify existing examples of data linkage being undertaken and assess the demand for further potential linkage projects.
3. Completing a privacy impact assessment (PIA) to identify the potential risks of a sample linkage project and determine if a PIA enables early identification of potential challenges.
4. Exploring the identified challenges, in combination with a review of the legislative and regulatory environment within which any future data linkage would occur, to assess how linkage can be facilitated through national infrastructure.

4.3. Research Design and Strategy

The research design structures the study to produce sufficient and appropriate evidence to address the research question as accurately, clearly, and unequivocally as possible (McGivern, 2006).

An observational rather than interventional or experimental approach was selected in order to obtain a clear representation of the current linkage environment and also to assess the attitudes and opinions of researchers working in the area. Observational studies such as this involve studying the research area without influencing, modifying or manipulating (Creswell, 2013). Observational studies tend to be more generalizable, faster and cheaper to conduct and can address a broader range of questions when compared to interventional studies, however they are more susceptible to bias and confounding, which must be addressed and controlled for through study design and advanced statistical methods (Katz, 2006).

There are three overarching approaches to conducting scientific research; qualitative, quantitative and mixed methods (Table 4.1). While qualitative research is usually focused on exploring and understanding a research topic, quantitative research is generally utilised to test objective theories by assessing relationships between variables. A mixed methods approach involves integrating both quantitative and qualitative data in order to obtain a more complete understanding of the research topic, while minimising the limitations of the two individual research approaches (Creswell, 2013).

Table 4.1: Quantitative, mixed and qualitative methods of data collection and analysis

Quantitative Methods	Mixed Methods	Qualitative Methods
Pre-determined methods	Both pre-determined and emerging methods	Emerging methods
Instrument based questions	Both open- and closed-ended questions	Open-ended questions
Performance data, attitude data, observational data, and census data	Multiple forms of data drawing on all possibilities	Interview data, observation data, document data, and audio-visual data
Statistical analysis	Statistical and text analysis	Text and image analysis
Statistical interpretation	Across databases interpretation	Themes, patterns interpretation

Source: Creswell (2013)

Mixed methods research is defined as "both a method and methodology for conducting research that involves collecting, analysing, and integrating quantitative and qualitative research in a single study or a longitudinal program of inquiry" (Creswell, 2008; pp.9). As demonstrated by Creswell (2008), a mixed methods approach can be used to review the research topic from multiple angles and its intuitive design means it can be used to obtain information on the "real life" situation. As the aim of this research was to explore the current research environment for linkage with longitudinal research data in Ireland, mixed methods was deemed the most appropriate approach.

Furthermore, a concurrent triangulation mixed method design (Figure 4.1) was selected in order to collect both qualitative and quantitative data concurrently (Creswell, 2013). By collecting both types of data, the two complementary data types can be used to address the research questions and the strengths of both forms of research are exploited. This would allow explanatory research on the existing linkage examples and future requirements but also allow for an exploratory review of unquantifiable elements such as researchers attitudes and perceptions in relation to the challenges and benefits of linkage.

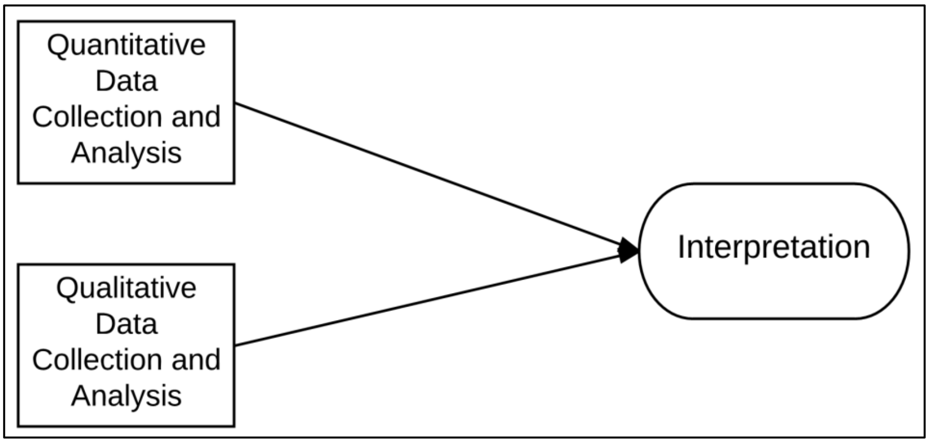

Quantitative Data Collection and Analysis
Qualitative Data Collection and Analysis
Interpretation

Figure 4.1: Concurrent triangulation mixed method design (Source: Creswell, 2008)

4.4. Research Methods

A combination of research methods were utilised to achieve the research aims outlined in Section 4.2.

4.4.1. Literature review

The initial stage of this research project involved conducting a review of existing relevant literature. As highlighted by DePoy and Gitlin (2015), conducting a literature review facilitates assessing the existing research and knowledge base of the topic of interest, determining how additional research will contribute to this existing body of knowledge and allow for the focusing and refining of the research question and strategy.

While the literature review was used to collect and collate evidence of the benefits and challenges of administrative data linkage, as recommended by Yin (2009), it was also used to develop and direct the later stages of the research. In particular, the evidence discovered during the literature review was used to inform the questions included in the questionnaire outlined in Section 4.4.2.

4.4.2. Survey of longitudinal researchers

A survey of longitudinal researchers was conducted to assess the current administrative data linkage environment in Ireland.

The aim of this survey was to:

1. identify the benefits and challenges of data linkage from the longitudinal researchers' perspective
2. identify existing examples of administrative data linkage and assess demand for further linkage projects
3. assess baseline knowledge among longitudinal researchers of key legislative and regulatory requirements

4.4.2.1. Survey instrument – questionnaire

A concurrent mixed methods approach was used for the primary data collection aspect of the research with data collected using a questionnaire. Within the design, there was unequal emphasis on the qualitative and quantitative data with higher focus on the quantitative.

The questionnaire included a combination of open- and closed-questions to facilitate the collection of both quantitative and qualitative data. While the closed-ended questions were used to collect data on topics such as demand for future data linkage, open-ended questions were used to collection information on participants' opinions and perceptions without biasing or guiding them with set answer options. Using a combination of the two question types enabled maximising the strengths of each question type, as outlined in Table 4.2, while also addressing the limitations of each.

Table 4.2: Strengths and limitations of closed- and open-ended questions

Closed-ended Questions	
Strengths	**Limitations**
1. Straightforward responses can be obtained 2. A large cohort can answer questions in a short period 3. Responses can be compared across groups 4. Statistical analysis can be conducted to describe and compare responses	1. The researcher is uncertain how respondents interpret or understand the questions 2. Issues relevant to respondents may not be captured 3. Respondent answers may reflect socially desirable responses
Open-ended Questions	
Strengths	**Limitations**
1. Highly sensitive issues can be explored 2. Nonverbal behaviours can be captured and analysed 3. Issues salient to respondents can be identified 4. Meaning of questions to respondents can be identified	1. Respondents may not want to address sensitive issues directly 2. Extensive time is required to collect information and analyse information 3. Responses across groups cannot be readily compared

Source: DePoy and Gitlin (2015)

The questionnaire was developed and refined using the Delphi method, which involves a process of iterative reviews and feedback by a panel of subject experts (Linstone and Turoff, 2002). Drafts of the questionnaire were reviewed by two leading longitudinal researchers in order to refine the questions and answer options and to obtain feedback and suggestions and updates to the questionnaire were made accordingly. The purpose of this Delphi process was to develop an effective measurement tool and also to reduce participant burden.

The questionnaire was developed using information obtained from the literature review, questions used in existing longitudinal study reviews and expert opinion. Questions on longitudinal research priorities were developed using input from a recent review of longitudinal studies in the United Kingdom, administered by the ESRC (Townsley, 2016). Additional questions, specific for the Irish research environment, were also included. These are based on recent work from key stakeholders such as the HRB (Moran, 2016) and HIQA (HIQA, 2017c). Options for potential future linkages were selected based on existing linkage projects in other countries. The included topics, questionnaire items and their related survey aims are detailed in Table 4.3. The complete set of questions included in the questionnaire is available in Appendix C.

Table 4.3: Survey aims and corresponding questions and topics

Survey Aim	Questionnaire Items	Topics Covered
Identify the benefits and challenges of data linkage from the longitudinal researchers' perspective	Questions 4, 4a, 9, 9a and 11	Perceived benefits; experienced challenges; potential administrative data linkage facilitators
Identify existing examples of administrative data linkage and assess demand for further linkage projects	Questions 5, 6, 7, 8, 10 and 10a	Successful and unsuccessful data linkage attempts; potential administrative datasets
Assess baseline knowledge among longitudinal researchers of key legislative and regulatory requirements	Questions 12, 13, 14, 15, 16, 17, 18, 19, 20 and 21	HIQA information standards; data protection; DASSL model; PIAs

4.4.2.2. *Questionnaire design and distribution*

The questionnaire was designed and administered online through Qualtrics software (www.qualtrics.com). An online survey distribution was selected as existing literature suggests this method is more economical and efficient compared to other methods such as face-to-face, written or telephone surveys (Van Selm and Jankowski, 2006). Other benefits such as the absence of interviewer bias, improved participant anonymity, the digitalisation of data during data collection removing the need for data entry and increased convenience for respondents have all been cited as benefits to online surveys (Sills and Song, 2002, McGivern, 2006).

Additionally, a key aspect of the questionnaire is the inclusion on open-ended questions and, as demonstrated by Sheehan (2001), respondents are more likely to respond to open-ended questions in online compared to paper based questionnaires. Furthermore, Paolo et al. (2000), concluded that respondents provide longer responses to open-ended questions in online surveys. The online questionnaire also utilises dynamic routing to ensure respondents are only asked relevant questions which would not have been possible with a paper based survey.

A review of online questionnaire design demonstrated that the use of radio buttons increased response rates and reduced missing data, whereas the use of text boxes improved the quality of responses (Couper et al., 2001). Therefore, a combination of these methods were include in the questionnaire in order to reduce missing data while maintaining data quality.

However, one disadvantage of online surveys is they are subject to lower response rates when compared to more traditional paper based surveys (Hohwü et al., 2013), which may have affected the response rate of this research

4.4.2.3. Respondent sampling

Selecting a sampling frame was identified as a major methodological issue for online surveys (Van Selm and Jankowski, 2006). The aim of the primary research was to recruit and survey researchers based in Ireland who conduct analysis using longitudinal survey data. However, due to the lack of a register of such researchers, a random sampling recruitment was not possible.

Non-probability sampling, which involves selecting potential respondents based on their convenience or availability, was therefore chosen (Creswell, 2013). Non-probability sampling in this way risks introducing potential biases into the study through sampling error (Weisberg, 2005). Sampling error occurs when only a subset of the population are sampled meaning all members of the population do not have equal chance of inclusion in the study and results in a study sample which does not reflect the overall population of interest (Weisberg, 2005). Also, using non-probability sampling, it is often not possible to quantify the risk and extent of sampling error as the probability of selection cannot be calculated due to a lack of information about the total study population (Weisberg, 2005, Van Selm and Jankowski, 2006). One recommended method to avoid sampling error in online surveys is to include the entire population of interest in the sampling frame (Sills and Song, 2002). Purposive sampling such as this can be implemented when the population of interest is small and unique as it allows a focus on a sample directly relevant to the research topic (Bryman, 2012). In an effort to achieve this, all current longitudinal studies were included in the survey recruitment.

Four established longitudinal studies which are currently active in Ireland where selected for inclusion in the sampling frame: Growing up in Ireland (GUI); The Irish Longitudinal Study on Ageing (TILDA), and the Intellectual Disability Supplement to The Irish Longitudinal Study on Ageing (IDS-TILDA) and the Maternal Health and Maternal Morbidity in Ireland (MAMMI) study. Attempts were made to contact researchers working on these projects either through publically available staff mailing lists, gatekeepers for the projects, or generic study contact emails.

A minimum sample size of 30 participants was sought. However, as analysis was restricted to descriptive statistics and no inference to the wider population were made, sample size calculations were not deemed necessary. No gender or age restrictions were applied to the recruitment, however, due to the recruitment being conducted through academic workplaces, no children under 18 years were enrolled.

4.4.2.4. Data management and analysis

Data from the questionnaire was collected using Qualtrics software (www.qualtrics.com) and exported into Stata 14 (StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.) for analysis. Cleaning and coding was performed on variables: numeric codes were applied to quantitative variables and thematic analysis was conducted on open-ended questions. As the aim of the survey was to collect descriptive statistics, detailed statistical analysis was not merited, particularly given the small sample size.

4.4.3. Privacy Impact Assessment

As identified in Chapter 3, conducting a PIA will become an essential part of any research project including data linkage projects. If both the administrative data custodians and the survey researchers qualify as data controllers then there will be a responsibility on both parties to ensure an appropriate PIA has been conducted in advance of any processing.

A PIA is defined as “a process which assists organisations in identifying and minimising the privacy risks of new projects or policies” (ICO, 2014; pp.5). The aim of a conducting a PIA is to identify privacy risks in advance, by reviewing the proposed uses of personal data, so the potential risks can be mitigated or eliminated before a project begins (HIQA, 2010). As presented in Figure 4.2, there are several stages which should be incorporated into a PIA in order to effectively identify and assess risks, develop measures to minimise the risks and incorporate these measures into the proposed project.

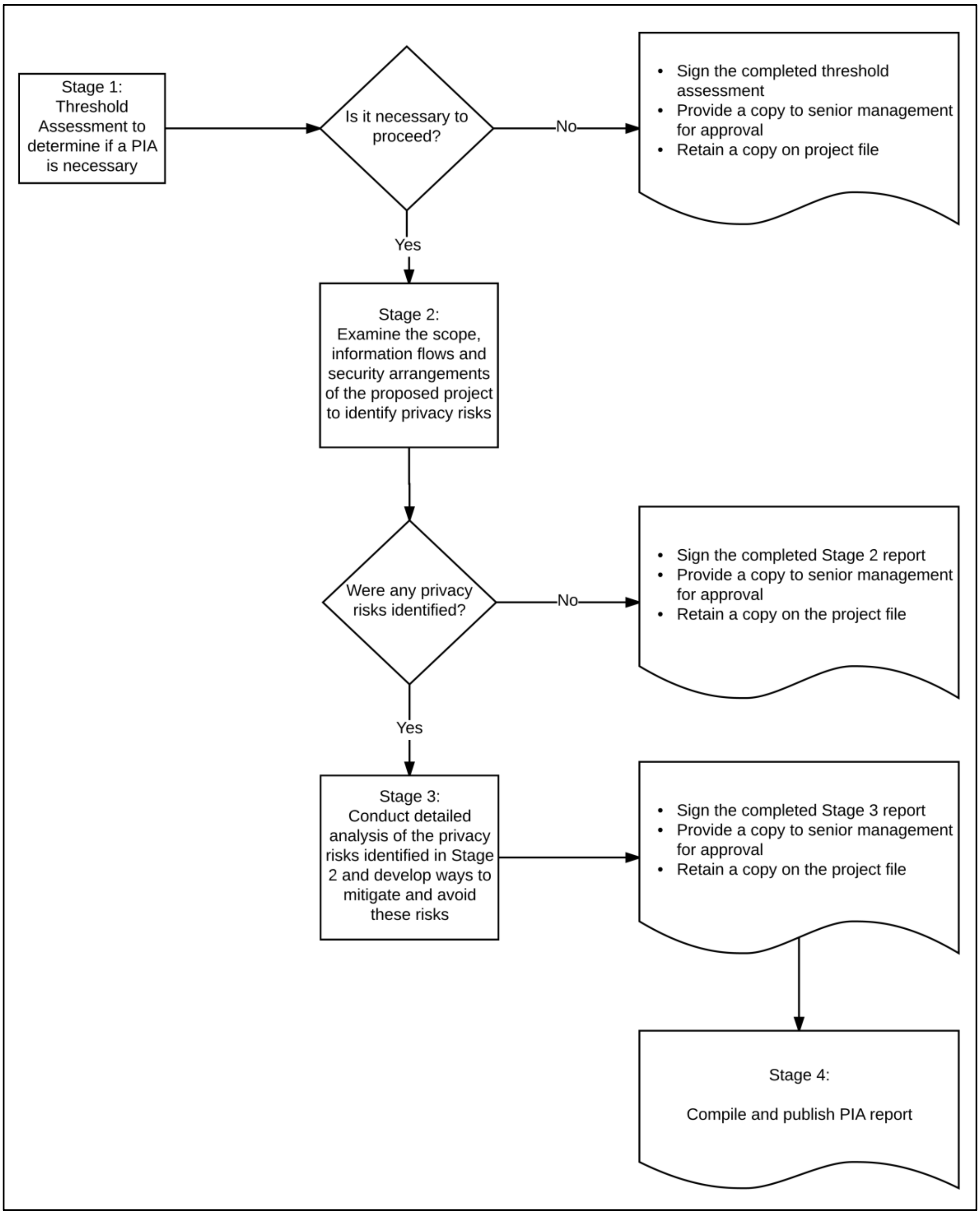

Stage 1: Threshold Assessment to determine if a PIA is necessary
Is it necessary to proceed?
No
Sign the completed threshold assessment
Provide a copy to senior management for approval
Retain a copy on project file
Yes
Stage 2: Examine the scope, information flows and security arrangements of the proposed project to identify privacy risks
Were any privacy risks identified?
No
Sign the completed Stage 2 report
Provide a copy to senior management for approval
Retain a copy on the project file
Yes
Stage 3: Conduct detailed analysis of the privacy risks identified in Stage 2 and develop ways to mitigate and avoid these risks
Sign the completed Stage 3 report
Provide a copy to senior management for approval
Retain a copy on the project file
Stage 4:
Compile and publish PIA report

Figure 4.2: Privacy impact assessment (PIA) process (Source: HIQA, 2010)

In order to assess how a PIA could be used to identify potential privacy risks of a linkage project, a sample PIA was conducted as part of this research. This sample PIA used the template developed by HIQA as well as input from toolkits and guidelines developed by the French National Commission on Informatics and Liberty (CNIL, 2015) and the UK's Information Commissioner's Office (ICO, 2014).

An administrative dataset linkage on which to conduct the sample PIA was selected from the responses provided during the survey of longitudinal researchers. This ensured that the selected example reflected a data source that is of interest to the research community. Responses to the survey highlighted the HIPE system as an administrative dataset that longitudinal researchers commonly reported could benefit their research aims. The output from this PIA is available in Appendix B.

4.5. Ethical and Data Protection Considerations

Ethical concerns of research involving human participants focus on three main elements; the rights of the research participants and the nature and scope of their involvement, the behaviours of the researchers and the ethics underlying the research aims and design procedures (DePoy and Gitlin, 2015).

In order to address any ethical concerns, all participants were fully informed about the purpose of the study in advance of beginning the questionnaire though a participant information leaflet (Appendix D) and a consent form (Appendix E). Information sheets were also provided to 'gatekeepers' when attempting to obtain contact details for a study's research team (Appendix D).

Only participants capable of providing informed consent were included. No contact information was collected during the questionnaire to maintain the anonymity of the participants. All data was collected, stored and processed in accordance with the Data Protection Acts 1988 and 2003.

Ethical approval for the research was received from the Research Ethics Committee of the School of Computer Science and Statistics, Trinity College Dublin (Appendix F).

4.6. Conclusion

This chapter provided an outline of how the aims of the research will be achieved, as well as detailing the motivation for selecting the methodology and study design. This included a reasoning for selecting a mixed method approach, in addition to details of the individual elements of the research project: literature review, survey and a sample PIA.

The following chapter will detail the results and analysis of the survey of longitudinal researchers, as well as reviewing these results in combination with the output of the sample PIA against the information identified during the literature review detailed in Chapters 2 and 3. The potential limitations of the selected research design will also be addressed in Chapter7.

5. Chapter 5: Results

5.1. Introduction

This chapter discusses the results of the research methods outlined in Chapter 4. The results of both the survey of longitudinal researchers and the PIA will be explored in detail.

5.2. Questionnaire Results

As detailed in Section 4.4.2, one component of the research methodology involved a questionnaire distributed to researchers working on longitudinal studies in Ireland. The main aim of this questionnaire was to review the current administrative data linkage environment in Ireland by:

1. identifying the benefits and challenges of data linkage from the longitudinal researchers' perspective
2. identifying existing examples of administrative data linkage and assessing demand for further linkage projects
3. assessing baseline knowledge among longitudinal researchers of key legislative and regulatory requirements

The results of the questionnaire in relation to each of these aims are discussed below. As the aim of the survey was to collect descriptive statistics, detailed statistical analysis was not conducted and would not have been statistically sound given the small sample size. The questions included in the questionnaire are provided in Appendix C.

5.2.1. Survey response rate

From the four longitudinal studies included in the survey, 37 researchers responded to the questionnaire. Due to the issues discussed in Section 4.4.2.3, it was not possible to calculate an accurate survey response rate as the number of researchers who received the questionnaire invitation was not known. Individual item response rates in the returned questionnaires were high across all questions ranging from 62% to 100% (Table 5.1). The lowest response rate was for question 11 (62%) which was an open-ended question.

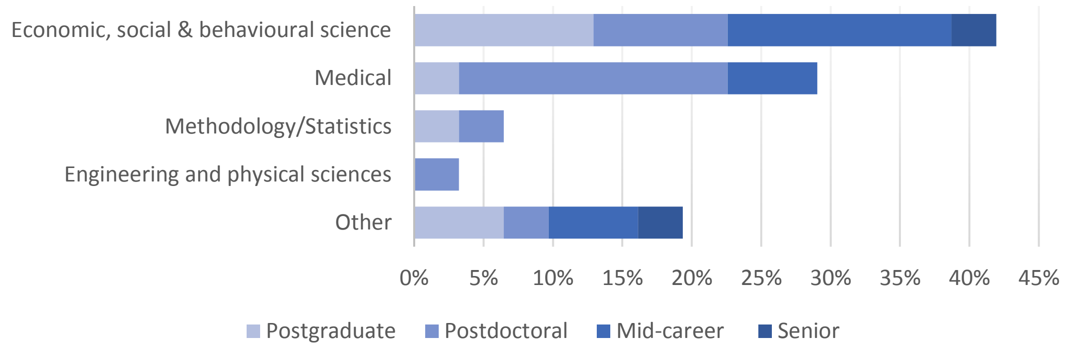
Table 5.1: Item response rates for questionnaire items

No.	Question	Baseline	Number of responses	Response rate
1	What is your career stage?	All	34	92%
2	What are the primary focus areas of your research?	All	33	89%
3	What is your primary research area?	All	31	84%
4	In your opinion, what are the potential benefits of administrative data linkage with research data?	All	31	84%
5	Have you ever attempted to link administrative data to your own research data or research data you have used?	All	32	86%
6	Were you able to successfully link the data?	*IF Q5 = YES*	12	100%
7	What administrative data did you link with your research data?	IF Q6 = YES	10	91%
8	What administrative data did you attempt to link to your research data?	IF Q6 = NO	1	100%
9	What barriers did you encounter during this linkage process?	*IF Q5 = YES*	10	83%
10	Would you see a benefit for your research to linking any of the following administrative datasets with your existing research data?	All	30	81%
11	What do you think would be the most important addition in Ireland to facilitate future administrative linkage projects in your research area?	All	23	62%
12	Are you aware of the DASSL model (Data Access, Storage, Sharing and Linkage) developed by the Health Research Board?	All	29	78%
13	Do you think the DASSL model would enable administrative data linkage in your research area if implemented?	*If Q12 = YES*	4	80%
14	Are you aware of the Information Management Standards for National Health and Social Care Data Collectors developed by the Health Information and Quality Authority (HIQA)?	All	29	78%

15	Have you ever conducted a privacy impact assessment (PIA) for your research?	All	29	78%
16	Have you ever used the HIQA privacy impact assessment tool?	*IF Q15 = YES*	1	100%
17	Did the HIQA PIA tool help to identify all challenges of data linkage in advance?	*IF Q16 = YES*	1	100%
18	Which barriers did the PIA not identify in advance?	*IF Q17 = NO*	N/A	N/A
19	Have you ever completed data protection training?	All	29	78%
20	Are you currently or have you ever been an Officer of Statistics as set out in the Statistics Act, 1993?	All	29	78%
21	Have you made any changes to your data collection or use policies in preparation for the incoming General Data Protection Regulation in 2018?	All	29	78%
22	Listed below are priority areas that longitudinal data could be used to address. Please indicate how important you think each of these longitudinal research areas will be in the future.	All	29	78%
23	Please indicate how important you feel each of these methodological or technological priority issues are in relation to longitudinal studies, thinking broadly across design, implementation and analysis.	All	29	78%

5.2.2. Survey respondent demographics

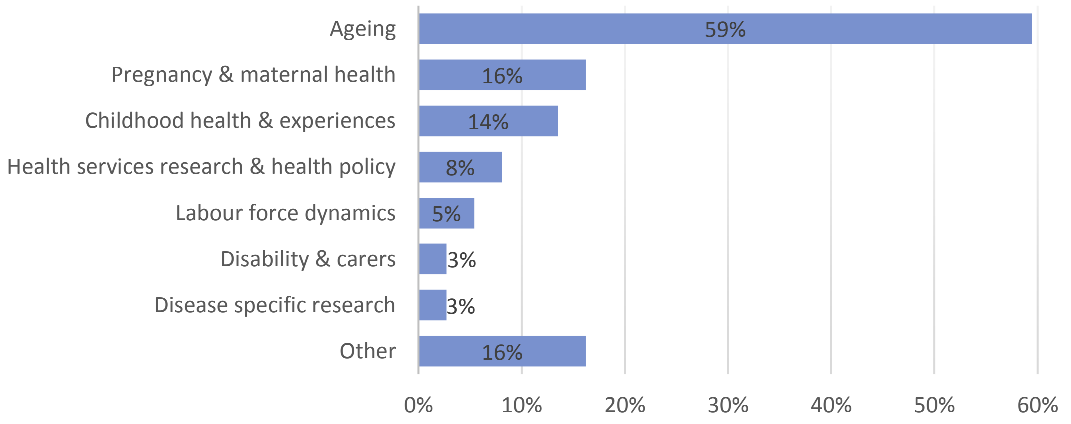
The career level and primary focus areas of the survey respondents are detailed in Figure 5.1. Of the 37 respondents, the majority were either postdoctoral (35%) or mid-career (32%) level. The highest number of respondents were from the economic, social and behaviour science disciplines (42%), followed by those whose primary research area was medical (29%). Of those respondents who reported other research areas (19%), disciplines such as psychology, midwifery and pharmacy were included.



Notes: values missing for 6 respondents

Figure 5.1: Primary research area and career level of survey respondents

In relation to the research focus of the respondents, the majority of survey respondents reported a single research focus (n=23; 70%) with lower numbers reporting two (n=8; 24%) and three or more focus areas (n=2; 6%). There was a strong preponderance for ageing research compared to the other research areas (Figure 5.2).



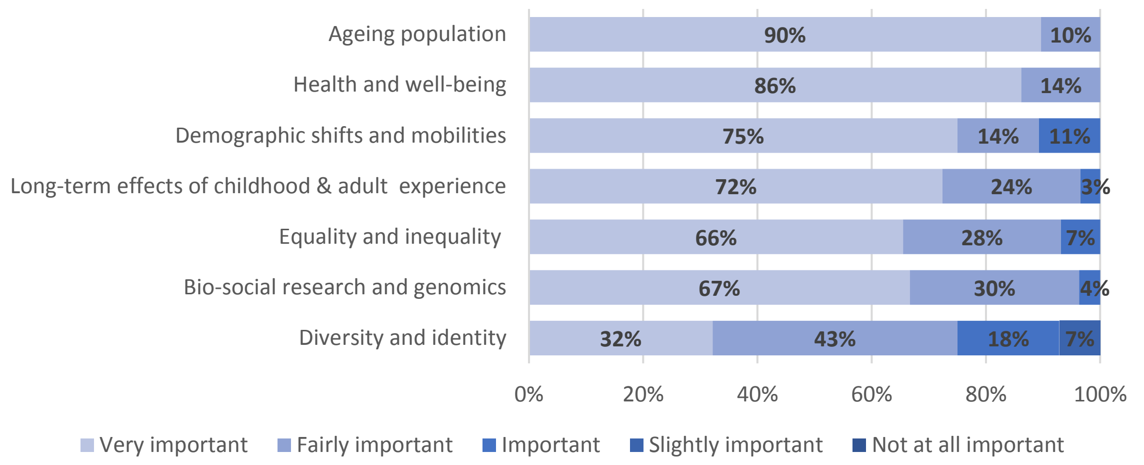
Notes: Values missing for 4 respondents

Figure 5.2: Research focus area of survey respondents

Survey respondents were also asked for their opinions on the future priority areas that longitudinal data can be used to address (Figure 5.3) and the methodological and technological priority issues for longitudinal studies (Figure 5.4).

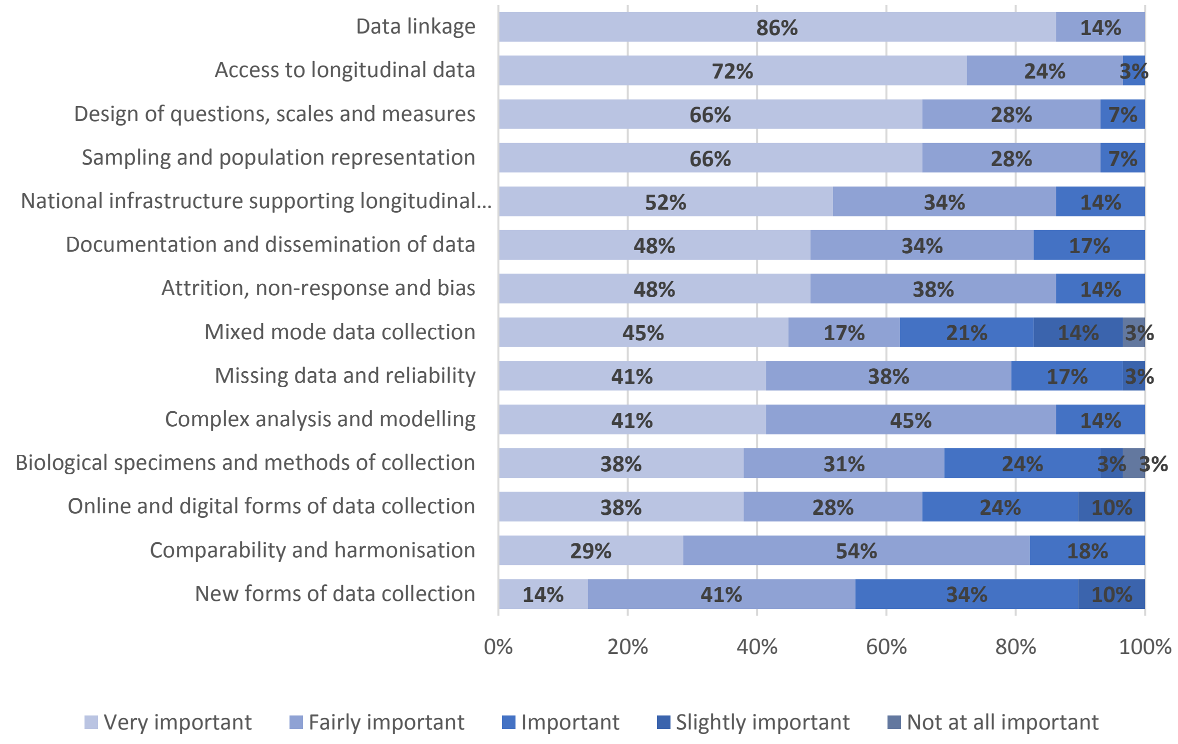
In relation to future priority research areas for longitudinal studies (Figure 5.3), a strong preponderance for ageing research was appreciated however this may reflect the majority of survey respondents who were already focused on ageing research (Figure 5.2).

Data linkage was reported as a priority for methodological and technological issues relating to longitudinal studies but this high weighting may have been influenced by the respondents being made aware of the aims of the survey through the study information sheet.



Notes: Values missing for 8 respondents

Figure 5.3: Future priority research areas for longitudinal studies



Notes: Values missing for 8 respondents

Figure 5.4: Methodological and technological priority issues for longitudinal studies

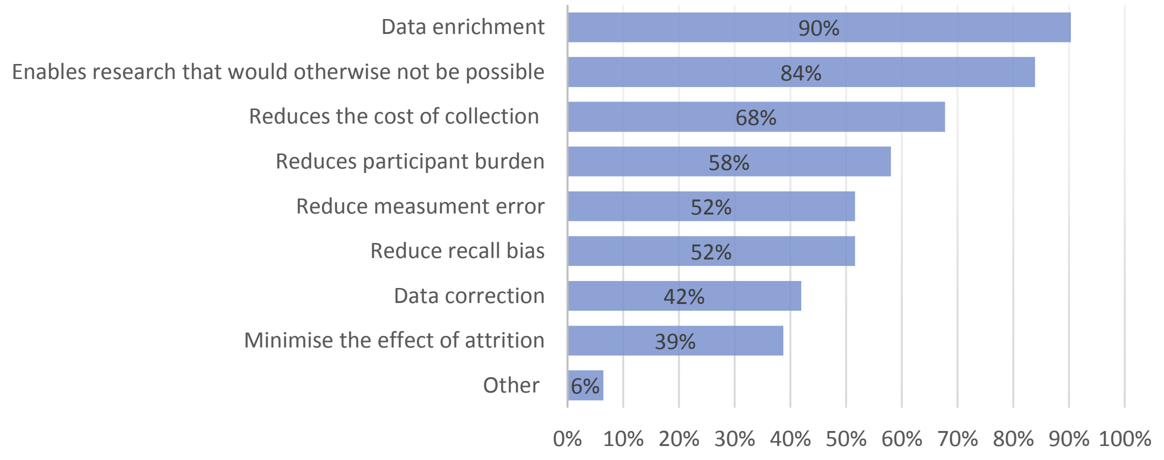
5.2.3. Survey Aim 1: Identify the benefits and challenges of data linkage from the longitudinal researchers' perspective

The aim of this section of the questionnaire was to elicit information on the perceived benefits of administrative data linkage from the longitudinal researchers' perspective. The questionnaire also sought to identify challenges to data linkage as experienced by the researchers.

There were 31 responses in relation to perceived benefits of data linkage, with respondents reporting between one and nine perceived benefits with an average of 5 benefits (Table 5.2). As presented in Figure 5.5, the most commonly reported benefit of administrative data linkage according to longitudinal researchers was the potential for data enrichment (90%) followed by the potential to enable research that would otherwise not be possible (84%). The majority of respondents also recognised the possibility of cost savings through data linkage (68%). A substantial proportion of the survey respondents also endorsed the use of data linkage to improve data quality with reducing measurement error (52%), recall error (52%) and data correction (42%) all reported by respondents. The potential to influence participation and response rates was also acknowledged, with 58% of the survey respondents reporting administrative data linkage could reduce participant burden and 39% reporting it could be used to minimise the effects of attrition. One respondent also included the potential to understand the study population characteristics as an additional benefit not contained within the set list.

Table 5.2: Reported number of perceived benefits of administrative data linkage

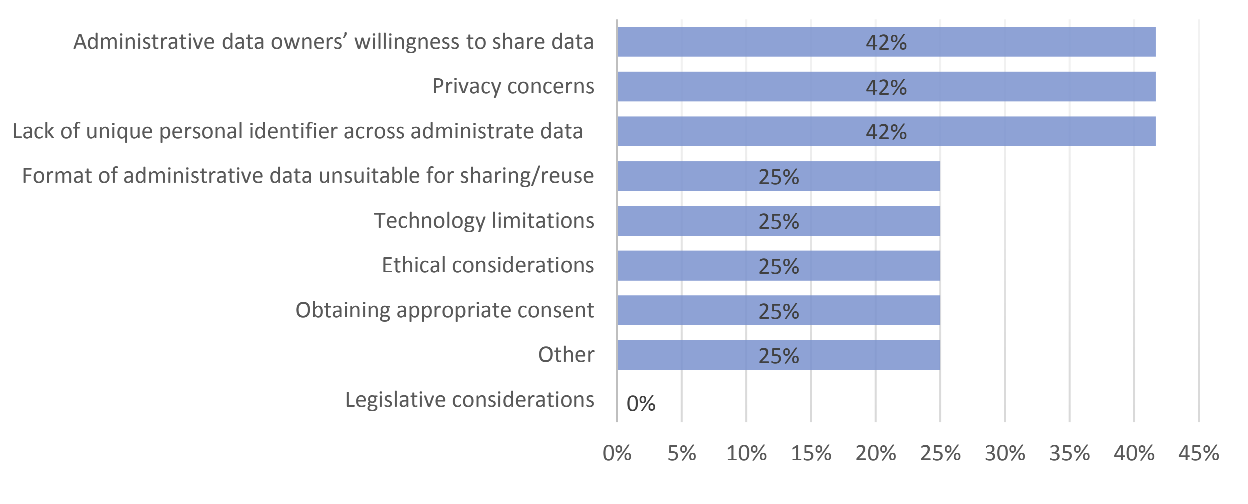
Count of perceived benefits	Number of respondents	% of respondents
1	2	6%
2	1	3%
3	6	19%
4	3	10%
5	7	23%
6	6	19%
7	3	10%
8	1	3%
9	2	6%
Notes: Values missing for 6 respondents		



Notes: Values missing for 6 respondents

Figure 5.5: Reported perceived benefits of administrative data linkage

There were 10 responses in relation to challenges associated with data linkage. Only respondents who had attempted administrative data linkage were asked this question, with respondents reporting an average of 3 challenges (range 2 -5). As presented in Figure 5.6, an equal proportion of respondents (42%) reported data custodian's willingness to share data, privacy concerns and a lack of a universal unique personal identifier as challenges when attempting administrative data linkage. A quarter of respondents reported unsuitable data formats, technology limitations, ethical considerations and obtaining appropriate consent as challenges to data linkage. Additionally, no survey respondents cited legislative concerns as a challenge for data linkage.



Notes: Values missing for 2 respondents

Figure 5.6: Challenges encountered when attempting administrative data linkage

When asked what would facilitate administrative data linkage in their research area, suggestions were provided by 23 respondents. This was an open-ended question allowing respondents to record details of any perceived facilitator of data linkage. The information provided in these answers were grouped into seven categories and then further grouped into four overall themes. These are detailed in Table 5.3 along with the frequency at which they occurred in the returned responses, however, it was possible for a respondent to provide details on multiple potential facilitators. Due to the volume of respondents reporting the inclusion of a UI across datasets to facilitate administrative data linkage in their research area this was included as a standalone theme. Additionally, the responses of two respondents indicated that they did not understand this question. However, this misunderstanding of questions is a recognised weakness of self-administered questionnaires (Ponto, 2015).

Table 5.3: Respondents' perceived facilitators of future administrative data linkage

Theme	Facilitator of future administrative data linkage	Number of respondents
Unique Identifiers (UI)	Presence of UI across datasets	11
Data Governance	Development of national administrative data bank	4
	Ethical framework for data linkage	3
	Reinforcing legislative approval for data linkage	2
Technology	Electronic health records	5
Data Quality	Standardised data collection and recording	1
	Identifying and filling data gaps	1

5.2.4. Survey Aim 2: Identify existing examples of administrative data linkage and assess demand for further linkage projects

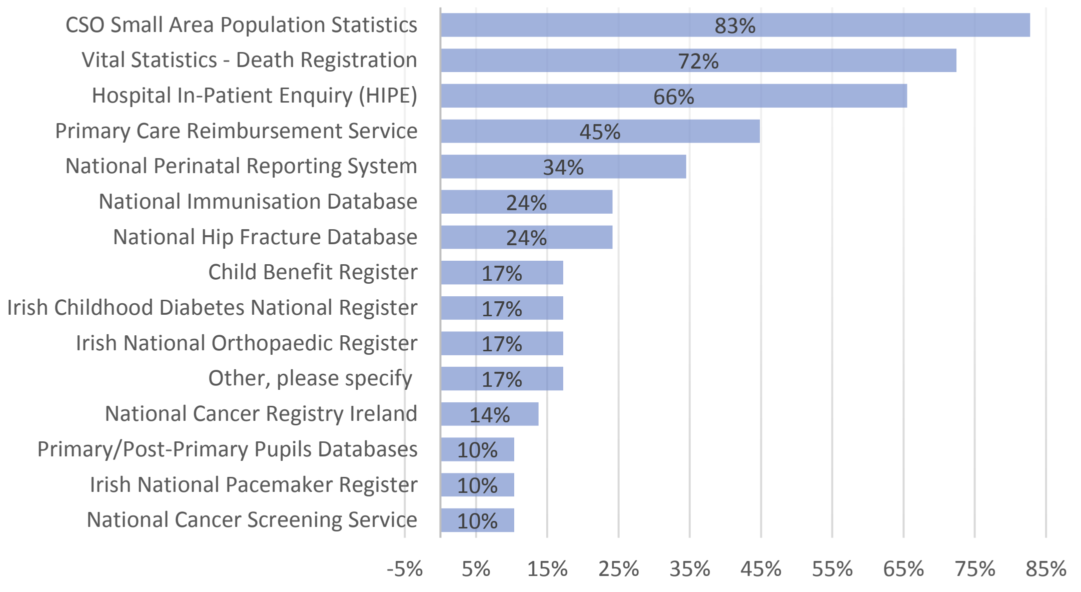
The aim of this section of the questionnaire was to determine if data linkage was occurring between longitudinal studies and administrative data sources and to assess the demand for further linkage.

Of the returned questionnaires, 12 respondents reported previously attempting to link administrative data to their longitudinal data and of these, all but one had been successful. The most common successful linkages (n=4) had involved the Central Statistics Office (CSO) small area population statistics which incorporate information on local areas such as population demographics, access to green spaces, broadband connectivity, noise and radon exposure and local infrastructure such as accessibility of healthcare facilities, food retail outlets and petrol stations. Two respondents reported linkage with local health records and another two with the Primary Care Reimbursement Service which contains information on medications obtained through the General Medical Services (GMS) card. Individual studies also reported successful linkage with Joint Replacement Registry Data, the National Psychiatric Inpatient Reporting System and the Pathway Accommodation and Services System. Remarkably, the respondent who reported unsuccessful data linkage was also attempting to conduct linkage with the CSO small area population statistics.

In relation to demand for further linkage, respondents were presented with a list of 14 potential datasets along with the option to suggest additional administrative data sources which could potentially benefit their longitudinal research. The survey respondents reported a potential research value of linkage with an average of four administrative datasets per researcher (range 1 - 9).

As presented in Figure 5.7, the most commonly requested dataset was the CSO small area population statistics (83%) followed by death registration data (72%) and HIPE data (66%). Additional suggested datasets included revenue data, unemployment data from the live register, utilities information such as drinking water quality, as well as healthcare related information such as GP records, infectious disease records held by the Health Protection Surveillance centre (HPSC) and the National Intellectual Disabilities Database (NIDD).

The administrative datasets perceived to have the highest research value were relatively standard across the research disciplines with the CSO small area population statistics, death registration, HIPE system and National Perinatal Reporting System appearing as principally requested administrative datasets for ageing, pregnancy and maternal health and childhood health (Table 5.4).



Notes: Values missing for 7 respondents

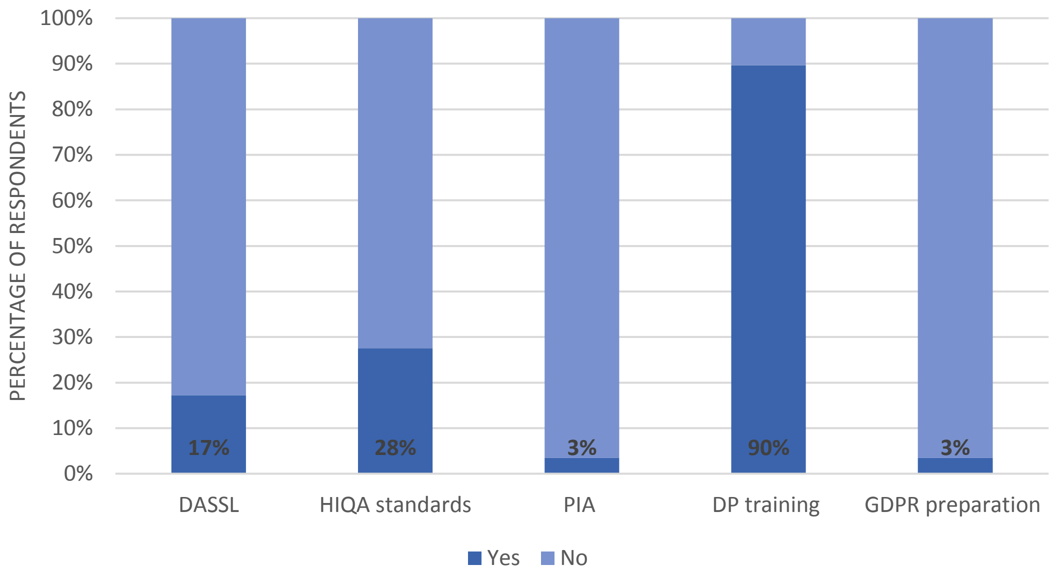
Figure 5.7: Datasets for which respondents reported a research benefit of linkage with research data

Table 5.4: Most requested administrative datasets by primary research focus (proportion of respondents requesting dataset)

	Primary research focus		
	Ageing (N=20)	**Pregnancy and maternal health (N=5)**	**Childhood health and experiences (N=5)**
1	CSO Small Area Population Statistics - (85%)	National Perinatal Reporting System - (100%)	CSO Small Area Population Statistics – (100%)
2	Vital Statistics - Death Registration - (85%)	Hospital In-Patient Enquiry (HIPE) - (80%)	National Perinatal Reporting System - (100%)
3	Hospital In-Patient Enquiry (HIPE) - (60%)	CSO Small Area Population Statistics - (60%)	Vital Statistics - Death Registration - (80%)

5.2.5. Survey Aim 3: Assess baseline knowledge among longitudinal researchers of key legislative and regulatory requirements

The final aim of the survey was to assess the baseline knowledge among longitudinal researchers of key legislative and regulatory requirements. The results of this review are presented in Figure 5.8.



Notes: DASSL – Data Access, Storage, Sharing and Linkage model; HIQA standards – Health Information and Quality Authority's Information Management Standards for National Health and Social Care Data Collectors; PIA – Privacy Impact Assessment; DP – Data Protection; GDPR – General Data Protection Regulation

Figure 5.8: Baseline knowledge of longitudinal researchers of key legislative and regulatory requirements

With the exception of data protection training, the proportion of respondents reporting awareness or implementation of key topics was very low. Only 17% of the researchers who responded to the survey were aware of the HRB DASSL model (Moran, 2016). However, of the 5 respondents who were aware, all reported the model would enable administrative data linkage in their research area. A slightly higher proportion of respondents (28%) were aware of the HIQA Information Management Standards but only 1 respondent had completed a PIA which is an essential component of these standards. While 90% of respondents reported completing some form of data protection training, there appears to be little preparation in advance of the introduction of the GDPR with only 1 respondent reporting a change in policies to account for the new legislation. Additionally, only 17% of the respondents reported being current Officers of Statistics which legally requires them to comply with data use and confidentiality requirements as set out in the Statistics Act, 1993 (CSO, 2014b).

5.3. Privacy Impact Assessment

As outlined in Chapter 3, conducting a PIA will become an essential part of any research project including data linkage projects. In order to assess how a PIA would influence a linkage project, a sample PIA was conducted using the template developed by HIQA as well as input from toolkits and guidelines developed by the French National Commission on Informatics and Liberty (CNIL, 2015) and the UK's Information Commissioner's Office (ICO, 2014). This sample PIA was conducted using a fictitious sample longitudinal study (SLS) and data from the HIPE system, the output of which is available in Appendix B.

Based on the Stage 1 threshold assessment of the sample PIA (Appendix B – Stage 1), a data linkage project between a longitudinal survey and the HIPE system meets several of the PIA criteria and therefore requires a full assessment. The result of this threshold assessment has implications for any study wishing to undertake an administrative data linkage project such as this, as it demonstrates that a PIA will need to be undertaken in advance in order to remain HIQA and GDPR compliant.

Based on the full sample impact assessment (Appendix B – Stage 2), five potential risk were identified which are presented in Table 5.5, along with an appraisal of their overall risk based on their likelihood of occurrence and their potential impact. A review of these five risks suggests they are likely to be relevant for all administrative data linkage projects as the issues are relatively universal. For example, any attempted data linkage project would need to address the issues identified in risk 3 and 4 in relation to what data will be shared and linked and how matches across datasets will be identified. Also, all data linkage projects will likely need to address the issues identified in risks 1, 2 and 5 relating to the confidentiality and approved uses of shared data and how the secondary use of data will be legally and ethically compliant.

Interestingly, the five risks identified in the PIA could all be successfully addressed by the introduction of control measures such as contracts and data transfer agreements (DTA) between the research study and HIPE, as well as preapproved consent forms and linkage protocols. (Appendix B – Stage 3). Also, the proposed solutions were effectively balanced against the aims of the linkage project while maintaining the privacy rights of the included participants.

However, while the five risks identified in this sample PIA are likely to relate to all administrative data linkage projects, there are additional potential risks of other projects which would not be identified in this sample PIA. Therefore, as each potential linkage project needs to be assessed on its own merits, the results of this PIA should not be considered as a comprehensive list of all potential risks.

Table 5.5: Privacy risk rating of risks identified in sample PIA using HIPE data

Risk	Likelihood	Impact	Risk Rating
1. This project constitutes a new use for data beyond that for which it was originally collected and this reuse of data must be legally compliant. There is a risk that the information booklet and consent form will not sufficiently notify participants about how their data will be used, meaning consent will not be fully informed. Additionally, the wording of the consent may not be sufficient to ensure HIPE can legally disclose personal health information to SLS.	Unlikely	Major	Medium
2. Both parties will have personal data disclosed to them, of which they are not the data controllers. There is a risk that information will be used inappropriately or disclosed further to additional parties.	Unlikely	Major	Medium
3. Not providing sufficient data from SLS to HIPE may result in inefficient matching and requesting insufficient data from HIPE to SLS may limit the research that can be conducted on the linked data. Both of these scenarios may require the linkage process be repeated which increases the opportunity for error or privacy risk.	Likely	Moderate	Medium
4. If an inadequate linkage process is conducted, there is a risk of incorrect matching which will result in incorrect information being assigned to participants. There is also a risk that details of participants who did not consent to inclusion in the project will be incorrectly sent to HIPE.	Likely	Moderate	Medium
5. Risks associated with responsibility in relation to data access requests. SLS will need to disclose the HIPE data to the participant which it relates to if a subject access request is received.	Unlikely	Minor	Low

One notable issue with the PIA is that it failed to identify and address challenges of data linkage which fall outside the remit of the privacy of study participants. For example, the quality and structure of administrative data was identified from the literature review as a major challenge to utilising data linkage but this was not identified as a risk in the sample PIA.

5.4. Conclusion

This chapter presented the results of the survey of longitudinal researchers and the sample PIA. The intention of the survey was not to collect data for detailed statistical analysis but rather to explore the current situation and the demand for further linkage projects and support services. There was a positive response to the survey with the aim of 30 returned questionnaires being surpassed. While the responses to the questionnaire established that researchers see a benefit to administrative data linkage and there are cases of it successfully occurring within some of the research studies, there is a prevailing demand from researchers to incorporate administrative data into their research analysis. The survey also indicated that researchers are aware of potential challenges to administrative data linkage but that baseline knowledge of key issues such as HIQA standards and the proposed DASSL model are low.

The sample PIA demonstrated that a full PIA would likely be required for the majority of administrative linkage projects with longitudinal studies, as the threshold assessment met several of the PIA criteria. The sample PIA identified five potential risks of the proposed project, all of which on review would relate to most projects involving sharing and linkage of data between a longitudinal study and an administrative data source. However, all the identified risks could be mitigated by incorporating interventions into the design of the study. While the sample PIA identified five universal risks, potential further risks for other administrative data linkage projects could not be assessed. Additionally, the sample PIA demonstrated that not all administrative data linkage challenges would be identified through the PIA and that issues such as data quality need to be addressed separately.

In the next chapter, the results of the survey of longitudinal researchers and the PIA will be explored and interpreted in combination will the evidence obtained from the literature review.

6. Chapter 6: Discussion

6.1. Introduction

Administrative data is defined as information collected and used as part of the routine day-to-day provision or management of public sector services and schemes including the healthcare system (MacFeely and Dunne, 2014). There is a rising awareness of the significant potential for research to maximise the use and output of this administrative data and international evidence has demonstrated it can be incorporated into research projects. However, Ireland currently lags behind other counties, many of which have established infrastructure for facilitating the use of administrative data in research projects for public benefit (Boyd et al., 2014).

The primary aim of this research was to identify *the benefits and challenges of linking health and administrative data with research data in Ireland*. Longitudinal research studies were selected as an appropriate example through which to examine these issues in the Irish context, given the presence of several high quality established longitudinal studies and evidence showing longitudinal studies are successfully utilising administrative data linkage in other countries.

This chapter will explore the results of the survey of longitudinal researchers and the PIA in combination with the evidence obtained from the literature review and their application to the aim of the project as detailed above.

6.2. Benefits of Administrative Data Linkage

The review of relevant literature demonstrated there are many benefits to administrative data linkage with longitudinal studies, which are summarised in Table 6.1. Interestingly, the evidence from the literature suggests these benefits are bidirectional with potential positive outcomes for both individual studies and the wider research community but also for the custodians of the administrative data.

All of the benefits identified in the literature were positively endorsed by the respondents to the longitudinal researcher survey, suggesting that, despite the under-utilisation of data linkage in Ireland, researchers are aware of the potential benefits of incorporating administrative data into their research (Table 6.1).

Table 6.1: Comparison of benefits of data linkage identified from literature and researcher survey

Theme identified in literature review	Corresponding survey responses (Question 4)	% of survey respondents endorsing the benefit
Data enrichment	- Data enrichment - Enables research that would otherwise not be possible	97
Data correction	- Data correction - Reduce measurement error - Reduce recall bias	74
Reduces cost of data collection	- Reduces cost of data collection	68
Reducing participant burden	- Reduces respondent burden	58
Sample maintenance and attrition	- Minimise the effect of attrition - Other*	39

* Other category added to sample maintenance and attrition as the free text response related to understanding the characteristics from which the study sample is drawn

The main benefit endorsed by the survey respondents related to data enrichment, which along with the potential to enable research which would otherwise not be possible were the two most frequently reported benefits.

This potential for widening the evidence base available to longitudinal researchers was also frequently cited in the literature particularly in relation to collecting data that would be difficult or impossible to obtain using traditional survey methods. There were also examples of using data linkage to increase the follow-up period of a study by supplementing with information on events which occurred outside the existing study period. Additionally, there is potential to use administrative data to incorporate more timely up-to-date data into research. This could be of increased benefit following the introduction of EHRs, meaning increasing levels of routine health data is collected digitally, facilitating faster extractions, sharing and linkage processes.

Data correction was similarly identified as a key benefit in both the literature and the survey responses. This is due to the unique potential ability of administrative data to address some of the inherent methodological weaknesses of longitudinal studies, namely item non-response, measurement error, particularly recall bias, and panel conditioning. This is a major incentive to implement data linkage as any analysis produced is dependent on the quality of the raw data so efforts for data correction are a justified use of resources.

In addition to this effective use of resources, data linkage also presents an opportunity to reduce the cost of collection, a benefit which was identified in the literature and also ranked highly by the survey respondents. In comparison to the costs associated with traditional survey data collection, administrative data represents a significantly cheaper alternative. While the literature did acknowledge a cost associated with administrative data extraction and cleaning, these were significantly offset by the saving in data collection. Additionally, the WADLS linkage system secured grant funding for research enabled by the system which represented a 10-fold return on investment (Holman et al., 2008). As the cost-benefits of using administrative data are amplified when records are digitised, this potential for return on investment could be used as justification for dedicating resources to developing EHRs in Ireland, as there is potential to recoup the costs through effective secondary use of the data.

The maintenance of a study's sample over the lifetime of the project in order to ensure results are generalizable to the population of interest was identified as a major issue in the literature review. As detailed in Section 2.4.4.4, the ability of a longitudinal study to maintain this representativeness is affected by attrition, particularly accumulative attrition over consecutive waves of collection which also limits the research potential due to smaller available sample sizes. Despite attrition being described in the literature as the 'Achilles' Heel' of longitudinal studies, the potential of administrative data to address the issue was only recognised by 39% of the survey respondents.

Two reasons for this disparity are hypothesised; firstly, the Irish longitudinal studies included in the survey are in their infancy when compared to the international literature meaning their samples are likely to still represent the baseline population. For example, TILDA has only collected data since 2009 (TILDA, 2017), whereas ELSA and HRS in comparison began in 2002 and 1992, respectively (Banks et al., 2016, HRS, 2016). For this reason, the samples being used by the survey respondents in Ireland may suffer less from attrition compared with the international studies, meaning researchers have not begun to fully investigate and understand possible solutions to attrition to the same extent as these studies which were identified in the literature review. Therefore, using administrative data linkage to address attrition may become more relevant for Irish researchers as the studies mature and continue over multiple waves.

Secondly, as the survey was aimed at individual researchers, the distribution of benefits of administrative data linkage likely reflect the focus of their individual research aims, such as enabling specific analysis, rather than the issues affecting the study as a whole. If the questionnaire had focused on Principal Investigators (PIs) or study management, overall study issues such as sample maintenance may have appeared higher on the ranking of potential benefits.

This lack of focus on the study-level issues may also explain the relatively low proportion of survey respondents (58%) reporting reducing participant burden as a potential benefit of administrative data linkage as minimising burden and attrition are intrinsically linked – long, difficult, embarrassing or uninteresting questionnaires are likely to results in higher rates of attrition at subsequent waves. The literature review demonstrated that administrative data can be used to facilitate analysis on sensitive data and reduce the amount of information that needs to be collected directly from participants by utilising data that already exists in administrative datasets. It can also allow the focus of the survey to move away from collection of routine data towards topics such as personality, perceptions and self-rated measures which are not collected in administrative datasets.

Overall the evidence from the literature established that it is the combination of administrative data and research data rather than an outright replacement of one for the other which yields the benefits of data linkage. This method ensures that the recognised benefits of both data sources are maximised.

6.3. Challenges of Administrative Data Linkage

Although there are many potential benefits of administrative data linkage as identified in Section 6.2, it constitutes a secondary use of data for which there are practical, cultural, legal and ethical challenges and implications which need to be addressed for the benefits of data linkage to be fully realised.

Several potential challenges were identified in the literature review which are summarised in Table 6.2. When reviewing the challenges, there was less concordance in those identified from the literature review and the survey responses (Table 6.2). While the themes still matched, the proportion of respondents endorsing each of the challenges was lower than the identified benefits (Table 6.1). However, the sample size was relatively small (n = 12) as only those who reported attempting data linkage were asked about challenges.

Additionally, the survey responses only reflect the opinions of the research community and it is possible that challenges from the perspective of the administrative data custodians may differ, however exploring their perspective was beyond the scope of this project.

Table 6.2: Comparison of challenges of data linkage identified from literature and researcher survey

Theme identified in literature review	Corresponding survey responses (Question 9)	% of survey respondents endorsing the barrier
Consent	- Obtaining appropriate consent	25
Unique identifiers	- Lack of unique personal identifier across administrative data	42
Data ownership and the role of data custodians	- Administrative data owners' willingness to share data - Other (X2)*	50
Quality and structure of administrative data	- Format of administrative data unsuitable for sharing/reuse	25
Privacy and trust	- Privacy concerns - Ethical considerations	42
Technology limitations	- Technology limitations - Other (X1)^{X}	33

* Other category contained 2 responses which could be related to the role of administrative data custodians

X Other category contained 1 response which could be related to the lack of matching technologies to ensure matches were accurate

One of the main themes to emerge from the literature was the multifaceted issues surrounding obtaining appropriate consent for data linkage. While the prevailing view was that consent constitutes the main way to conduct legally and ethically sound data linkage, there was also sufficient evidence to demonstrate that depending on consent can negatively influence or limit the research potential of data linkage. The negative impact of consent dependence can occur in several ways including affecting the representativeness of the sample, introducing bias and limiting research in minority groups or on conditions which affect the ability to provide consent such as dementia. This issue of consent is further complicated as the factors influencing consent are not fully understood with the studies reporting varying and conflicting results.

Given the high importance placed on consent, both ethically and legally, it was unsurprising that consent was one of the most frequent issues that emerged during the literature review. However, despite the dominance of consent issues identified in the literature, this was not reflected in the views of the longitudinal researchers, with only 25% of survey respondents reporting difficulties with obtaining appropriate consent for administrative data linkage (Table 6.2). This also contrasted with the evidence of the sample PIA which determined consent was required to legally disclose the data for secondary use. There may be an issue with how the survey respondents classify 'appropriate' consent, and while this research did not allow for an assessment of the consent obtained for existing data linkage, the high level of data protection training (90%) among the sample indicates they should have a sufficient understanding of what level of consent is required for data linkage.

While the literature did establish methods to conduct data linkage without participant consent, these methods to sidestep consent, usually through ethical waivers, required lengthy and resource-heavy approval processes. These methods should be used with caution as even when consent is not legally or ethically required, evidence suggests that participants' preference is to still be consulted. Audrey et al. (2016a) demonstrated that participants want to maintain ownership and control over how their data is used, particularly when data linkage is used to investigate perceived socially sensitive issues. Through a qualitative analysis, authors demonstrated that while longitudinal study participants were not concerned about proposed linkage projects, the majority still wanted to be consulted in advance. This preference for prior consultation and consent increased when asked about more sensitive linkage topics, and the authors concluded that support of linkage was dependent on the stigma associated with the subject matter. This requirement to review each proposed linkage in isolation and that consent to one form of administrative data linkage does not imply a participant is open to broader linkage is reflected in the sample PIA conducted as part of this research, as it demonstrated that the PIAs are project specific.

Also, it should not be assumed that circumventing the consent process will remove the bias associated with data linkage. Even if the risk of consent bias as described above could be mitigated by bypassing the consent process through legal means, there remains an inherent bias in using administrative records for research. This is because only those accessing services will be contained within the administrative records and therefore included in the resulting research analysis. For example, in the case of healthcare records, only those unwell enough to access healthcare will be included and younger, healthier people or those with milder conditions may be underrepresented.

While this bias can be adjusted for it requires complex statistical weighting measures. However, this bias may not impact some research projects, particularly those focused on a specific disease or condition. For instance, if a research project aims to examine the healthcare resources used to treat a specific disease then analysing only those who access services may be sufficient.

One of the challenges which emerged from both the literature review and the survey of researchers was administrative data custodians' willingness to engage with potential data linkage projects. Administrative data custodians are legally responsible for protecting the privacy of their data subjects, however, evidence from the literature review suggests some of the reluctance to share and link data is due to fear and uncertainty of what is legally permissible. Additionally, other organisational issues such as data hoarding or efforts to protect resource investment is preventing custodians from engaging with researchers. These issues suggest there is a lack of clear data governance infrastructure or guidance at a national level.

Even when data is made available, issues surrounding data quality and structure lead to underutilisation and misuse. One of the main problems identified from the literature was that administrative data is not structured for research and accompanying explanatory documents such as data dictionaries are lacking. The introduction of the HIQA information standards and the development of the catalogue of national health and social data collections (HIQA, 2014a) may address these issues as it calls for standardised documentation across all datasets. The need for standardised collection and recording systems such as agreed terminology was also recognised as an issue with existing administrative data. However, the development of EHRs across the health system has the potential to address this, as enforcing standard coding systems and minimum datasets is easier to implement in digital records compared to paper charts.

Positive moves towards adopting these standards can be seen in the recent purchase by the HSE of a SNOMED clinical health terminology licence which had been recommend by HIQA (HIQA, 2014c).

Similarly, there was concordance between the examined literature and the survey responses in relation to UIs, with both suggesting the lack of UIs as a barrier to data linkage. Furthermore, incorporating UIs into administrative datasets was identified from the survey as the most frequently reported requirement to facilitate future linkage projects (Table 5.3). This is an interesting finding in light of the recent signing by the Minister for Health of the commencement order allowing for the use of the national IHIs across the healthcare system (HSE, 2017c). While the IHI has the potential to enable effective matching across datasets it can also be utilised to identify and remove duplications within datasets. However, it is important to acknowledge that evidence from the literature demonstrates that successful linkage is possible without UIs and that even using them will not ensure completely accurate matching as inaccuracies are still possible through recording, transcribing and linkage protocol errors.

One of the most noticeable concerns to emerge from the longitudinal researcher survey was that none of the respondents reported legislation concerns as a barrier to data linkage. This contrasts sharply with the proportion of respondents citing privacy, ethics and consent challenges, indicating that respondents are unaware that privacy and consent are enshrined in legislation and that a breach in privacy actually constitutes a breach in data protection legislation. This is concerning considering 90% of the sample indicated they had received data protection training. As the legal protection and implications surrounding data processing will be amplified in the GDPR, it is vital that researchers understand their legal obligations when undertaking data linkage. The high level of respondents reporting data protection training contrasts with the only 3% of respondents reporting changes in their data collection and use policies in advance of GDPR's introduction which again implies a lack of awareness of legal obligations.

Despite the challenges discussed above it is possible to overcome these challenges as successful data linkage has occurred both internationally, and to a smaller extent, in Ireland. It is vital to understand how further linkage can be facilitated to ensure the benefits are fully exploited.

6.4. How can Data Linkage be Facilitated in Ireland?

One of the aims of this research project was to review how administrative data linkage can be facilitated in Ireland. Survey respondents expressed high levels of interest in incorporating administrative data into their research and reported many perceived benefits associated this linkage. As highlighted by the DOHC in 2001, the secondary use of data requires achieving an appropriate balance between maximising the benefit of collected data and protecting people's right to privacy and confidentiality (DOHC, 2001). This focus on a balance between these competing entities of research benefit and participant privacy was repeatedly seen in the literature.

6.4.1. Longitudinal researchers' demand for data linkage

This research project aimed to determine if there was a demand for administrative data linkage among the research community in Ireland as the presence of such a demand would support the allocation of resources towards linkage facilitates and infrastructure, particularly at a national level.

The results of the survey of researchers demonstrates there is a strong demand to incorporate administrative data into research projects with the respondents indicating a desire to link with an average of four separate datasets per researcher.

Interestingly, one of the most frequently requested datasets, the CSO small area population statistics, which was requested by 83% of respondents, is not personally identifiable data as it reports at a geographical level ranging from the entire state to small local areas that typically contain between 50 and 200 dwellings (CSO, 2014a). The data is not unique for each individual as neighbours within the same small area will all have the same characteristics at the small area level. As this data is not identifiable it would eliminate issues of data protection. Also, a vast amount of this data is freely available for download from the CSO website. One issue with linking this administrative data to research data is that information on the location of participants is required. For example, researchers would need to have the participants' electoral district (ED) or address geocode. A linkage process is then required to associate the statistics for each CSO area to the relevant participant. Completing a PIA in advance of attempting a linkage project such as this would ensure researchers were aware of the requirement to collect the CSO area or geocode of each participant.

Of those survey respondents who reported successful data linkage, CSO small area population statistics was frequently included (30%) meaning it is possible to overcome these technical difficulties. However, in contrast to this positive outcome for some researchers, the single dataset reported as an unsuccessful data linkage attempt was also the small area population statistics. While it was beyond the remit of this research to explore the cause of the different outcomes of the attempts to link with the CSO data, it does suggest there may be inconsistencies in how data linkage is approved for individual projects.

Similarly, data from death registration, which was the second most frequently requested dataset among survey respondents (72%), is not subject to data protection legislation as it is restricted to living identifiable individuals. However, while exempt from data protection there are still likely to be ethical and information standards that must be addressed when attempting this linkage.

The third most requested dataset (HIPE) does however fall within the remit of data protection legislations. Therefore, any attempt to link with HIPE data, which was requested by 66% of the survey respondents, would be subject to strict control to protect patient's privacy. However, the sample PIA, which utilised HIPE to explore potential risks surrounding a linkage project demonstrated that it would be possible for a longitudinal study to achieve linkage with this administrative data.

The survey showed that researchers requested an average of four administrative datasets. If linkage was attempted in the current environment, the absence of a national infrastructure means this would likely require the researcher to coordinate with four separate administrative data custodians to achieve successful linkage. As shown by the sample PIA, a legal agreement is usually required between both parties meaning four different contracts would be developed and agreed. Also, to achieve the linkage, the researcher would have to send the identifiers of their participants to each of the four custodians, increasing the amount of data sharing required. This is coupled with the possibility that, without a standard IHI, the custodians may all use different identifiers increasing the complexity of any sharing and linkage attempts. For instance, linkage with the Primary Care Reimbursement Service (PCRS) would require the General Medical Services (GMS) medical card number, HIPE would require name, address and date of birth, and the Child Benefit Register would require Personal Public Service (PPS) Number.

Qualitative work by Moran (2016) has shown that this is the reality for researchers attempting to engage in data linkage in Ireland. Evidence from the literature demonstrated that this leads to a lack of clear policies and varying levels of legal and ethical compliance (Stone et al., 2005, Brett and Deary, 2014, Hagger-Johnson, 2015). It is clear that, based on the complexities described above, coupled with the inconsistencies in current linkage practices, to meet the demand of researchers for administrative linkages, a national approach is required. This would ensure the benefits of data linkage are maximised while addressing the challenges consistently.

6.4.2. Reviewing a national infrastructure for administrative data linkage

As the DASSL model (Moran, 2016) is currently the only proposed infrastructure within the Irish context that could facilitate data linkage between administrative and research sources, the evidence developed and summarised in this research will be assessed against the proposed DASSL model as described in Section 3.4.

6.4.2.1. Information Governance

The proposed DASSSL mode incorporates many of the survey respondents' suggestions about what would facilitate data linkage in Ireland. Of the facilitators identified, issues such as the development of an ethical framework for data linkage and reinforcing the legislative framework for linkage would be housed within the governance element of the DASSL model. Having these facilities at a national level and ensuring that all proposed projects are routed through this infrastructure will ensure that the inconsistencies discussed in Section 6.4.1 are addressed, as standard linkage approval processes will be established under what Moran describes as "safe, effective and proportionate governance" (2016; pp. 39). Having a national infrastructure would also reduce the number of extensive contracts that would be required for individual linkages as identified through the sample PIA.

Additionally, a national infrastructure, such as DASSL, would also have an advantage over the current system of individually coordinated linkage projects in relation to data protection, ethics and standards. As new mandatory guidelines are introduced, the governance structure proposed by DASSL will be able to ensure only legally compliant projects are permitted. This would ensure a standardised approach to linkage projects compared to the current haphazard approach, as seen with the CSO small area population statistics linkage attempts reported as both successful and unsuccessful by survey respondents. The model also proposes linkage with the Data Protection Commissioner meaning there will be subject experts available to the governance review panel to ensure there are no ambiguities in how new legislation such as the GDPR are implemented. This would also assist with custodians' concerns about violating the legislation which was cited as a key cause of the reluctance to share data for linkage projects. Additionally, the low levels of GDPR preparation (3%) and awareness of the new HIQA standards (28%) among survey respondents would support the argument for addressing and enforcing these issues at a national level.

6.4.2.2. Technical linkage issues

Additionally, several of the survey respondents endorsed the establishment of a national administrative databank which is akin to the health research data hub of DASSL. The purpose of the data hub is to coordinate with the administrative custodians to collate the available data so that it can be used for research. The proposed data hub is based on international models such as the Secure Anonymised Information Linkage (SAIL) databank in Wales which contains individual-level administrative records relating to health and well-being and has successfully enabled high-quality research since its establishment (Ford et al., 2009). Creating a national databank such as this would reduce the duplication and complexity of multiple agreements and contracts as described in Section 6.4.1, as a single project application could be used to access data from several administrative sources.

While some of the other suggested facilitators, such as UIs and EHRs are not specifically referred to in the model, the establishment of a national linkage service within DASSL would maximise the use of these, once they are both implemented by the HSE and have become more widely used across the health service. Evidence from the literature review, confirmed that many of the international data linkage models, such as the Australian WADLS, hailed the presence of an UI across their administrative datasets as key to their success. This level of success is evident, not only in the high return on investment reported by WADLS, but also the high proportion of publications identified from the literature search that were based on data from WADLS (Appendix G).

6.4.2.3. Cultural issues

As highlighted by Jones at al. (2006), it is often not technological issues which are preventing wider utilisation of data linkage, rather a culture which does not promote, or in some cases stifles it, and the culture of sharing health data is reported to be particularly 'closed' in Ireland (Moran, 2016). This can be seen in the number of papers that cited data custodians' unwillingness to engage with researchers and the results of the survey of researchers in which it was reported as one of the leading challenges experienced. Fortunately, these cultural issues are acknowledged and addressed in the DASSL model.

DASSL aims to create a culture of data sharing and linkage which is promoted by professional bodies, educators and research funding bodies. Importantly, the DASSL model will remove the need for data custodians and researchers to be solely responsible for making decisions on the appropriateness of a proposed project, allowing instead for the administrative data to be added to a data hub that depends on the governance structures to ensure it is shared and linked appropriately.

The model also proposes incentivising both researchers and data custodians to, not just share their data, but to also make it available in reusable formats with comprehensive metadata and user guides. If this new ethos can be established, it would address the concerns which emerged from the literature review and the survey in relation to the quality and structure of administrative records.

Another key component of DASSL which will help address the current culture of data hoarding is active stakeholder engagement. Evidence from the literature and also from the recommendations that emerged from the sample PIA, show that it is necessary to involve all parties early in the research planning process. This is necessary to ensure the correct data is being requested and shared to achieve the research aims, that the provided data is being interpreted correctly, and that all parties are aware of their responsibilities. Significantly, the development of the DASSL model involved extensive shareholder involvement to ensure the proposed model met the needs of those who would be using it.

6.4.2.4. Limitations

As detailed above, the proposed DASSL model has the potential to address many of the challenges to administrative data linkage that were identified from the literature, researcher survey and sample PIA. It appears, based on this research, that DASSL represents an effective model for facilitating data linkage in Ireland. However, despite this clear potential to maximise the utilisation of administrative data in research, the model also has some distinct limitations.

Firstly, examples from other national linkage models have shown that, even when successfully established, there are still issues with data custodians' willingness to engage with the system, and the reluctance to share data with individual research projects is replicated when they are requested to add data to a national data hub. For WADLS, which is widely regarded as a prominent successful linkage system, it took over seven years to integrate some administrative data sectors and even then, the original custodians retained ownership and continued to control access and use of the data (Holman et al., 2008). If the DASSL model is to avoid pitfalls such as this, it will require substantial inter-agency collaboration so that organisational and collaborative requirements can be agreed. Considering this clear requirement for early stakeholder engagement, it is concerning that only 17% of the survey respondents reported an awareness of the model. To fully implement and utilise DASSL, there needs to be further efforts to raise awareness among the research community as their support, and use of the system once introduced, will be key to its success.

Additionally, there will be a need for data protection legislation to specifically address the activities of DASSL. While the incoming GDPR will restrict the sharing of sensitive personally identifiable data, as detailed in Section 3.2.3 it does allow for the development of national legalisation, particularly to address research and public health needs through Article 89. It is vital that any legislation introduced to facilitate DASSL is not too restrictive to an extent that it would prevent any potential future work of the linkage system. Depending on narrow legislation for specific purposes is inefficient and cumbersome, such as the current Infectious Disease Regulation (1981), which must be redrafted and republished each time a disease is added to the notifiable list. Instead it would be favourable to develop wide-ranging legislation which addresses the pathways of incorporating administrative data into DASSL rather than a list of specific datasets.

Furthermore, while the DASSL model does incorporate a strong governance structure and a public engagement component, Audrey et al. (2016a) demonstrated that public opinions of data linkage are complex and diverse and that successfully accommodating them into a governance structure that is satisfactory for all, or even the majority, of the included participants is incredibly difficult. Resolving the competing ends of the public retaining ownership over their data, implementing a national system which respects the rights of the population while maximising the potential research benefits, with its consequential public benefit, will be a major challenge for the linkage model.

Due to these difficulties and the vast amount of cultural, organisational, legislative and technical changes which are required, it is unsurprising that Brett and Deary (2014) concluded that successful linkage systems at a national level are rare. Therefore, it may be worthwhile introducing additional measures to facilitate data linkage which will not be affected by a delay in trying to establish a national system like DASSL.

6.4.3. Additional potential facilitators of data linkage

In the absence of a national linkage structure, PIAs are essential to ensure risks to study participants' privacy and confidentiality are minimised. Conducting a PIA in advance of any linkage project will assist in safeguarding both the data custodians and the researchers. The aim of a PIA is not to completely eradicate the risk to privacy at the expense of the aims of the proposed project, but to reach an effective balance where all potential risks have been identified and minimised. The PIAs will also assist the researchers, as they can be used in the planning stage of a proposed linkage project to ensure the required data is identified and linkage protocols are agreed.

Also, without the governance structure of the DASSL model, completing PIAs will act as indication of data protection compliance, fulfilling the GDPR's requirements to demonstrate accountability in relation to the principles of the legislation. This will become particularly important if a data breach occurs as it establishes that the parties involved acted appropriately and that risks were reviewed and addressed (HIQA, 2010).

The legal requirement for PIAs is in sharp contrast with the current level of PIA experience in the research community as demonstrated by the survey results, with just 3% of respondents reporting completing PIAs or incorporating them into research planning. This is coupled with the sample PIA showing a clear benefit for incorporating them into the planning stage of a research project. To ensure researchers attempting data linkage remain compliant with both GDPR and HIQA standards, awareness campaigns and training sessions should be developed so researchers have the necessary skills to complete PIAs and understand the wider implications of their legal responsibilities. However, as detailed in Section 5.3, the PIA does not address all the risks associated with administrative data linkage, and therefore there is a need to ensure that a completed PIA is not interpreted as a roadmap to a successful linkage project as issues such as data quality will not be incorporated.

Additionally, the current development of national EHRs presents a unique opportunity to positively impact how digitalised records are utilised in research. The additional of an 'opt in' which allows patients to consent to their information being used for research purpose is recommended to increase the amount of data available for research as well as reducing the burden on patients as they do not need to be contacted separately for consent or identified through registries for disease specific research (Kukafka et al., 2007, Willison, 2009, Sullivan et al., 2016). The level of consent can be modifiable, allowing patients to consent to certain forms of research or allow different levels of access. This is akin to the dynamic consent defined by Kaye et al. (2015) which uses interactive technology to enable a personalised interface that allows participants to view, alter and withdrawn their consent preferences in real time. This would enable greater access to primary healthcare data at a national level but strict information governance structures would be essential.

6.5. Conclusion

This chapter explored the results of the survey of longitudinal researchers and the PIA in combination with the evidence obtained from the literature. It also demonstrated how the evidence presented in this research supports the establishment of a national linkage infrastructure such as the DASSL model proposed by the HRB.

The final chapter of this research will conclude the key findings of the results and analysis for administrative data linkage in Ireland, address the limitations of the current research and also the implications for future research.

7. Chapter 7: Conclusion

7.1. Introduction

This research was instigated after noting a lack of administrative data linkage for research purposes in Ireland compared to other countries such as Australia. As this inconsistency had originally been identified though comparing the data coverage of longitudinal studies internationally, this was selected as an appropriate example to examine the Irish context, particularly given the presence of several high quality established longitudinal studies.

The aim of the research was to identify the benefits and challenges of linking health and administrative data and explore how this linkage could be facilitated further in Ireland. To achieve these aims a combination of a literature review, an appraisal of the legislative and regulatory research environment, a survey of relevant researchers and a sample PIA were utilised. The key findings from this body of work are presented below, followed by a review of the limitations of the research and the implications for future practice and research.

7.2. Key Findings

The evidence presented in this research has both practical and conjectural significance for administrative data linkage. The key findings which are outlined below can be used as a framework to direct the development of data linkage in Ireland.

The literature review identified several benefits of administrative data linkage, such as data correction, data enrichment and reduced costs of collection, which are being exploited by international researchers. Ultimately, the evidence shows that there are strengths and weakness of both data sources and that it is the combination of the two that enables research that otherwise may be impossible to achieve. Despite a relatively low number of linkage projects in Ireland, evidence from the survey demonstrates that there is an awareness of these potential benefits among longitudinal researchers.

However, administrative data linkage constitutes a secondary use of data for which there are practical, cultural, legal and ethical challenges and implications which need to be addressed for the benefits of data linkage to be fully realised. The challenges of data linkage identified though the literature were broadly similar to those recognised by the survey respondents, with the exception of legislation concerns which, despite encompassing many of the other challenges, was not identified by any of the survey respondents as a barrier to linkage. This reflects a possible lack of understanding of data protection requirements.

One of the prominent challenges to emerge from this research is informed consent, both the difficulties obtaining it and the implications of making data linkage dependent on it. If data linkage in Ireland is to be contingent on consent, then the factors influencing propensity to consent among the Irish population need to be explored. However, if data linkage is to proceed without seeking informed consent, there is a need to ensure the processes are fully compliant with legislation, standards and ethics. It is also important to consider participants' preferences, as the evidence presented here establishes that even when consent is not legally required, participants prefer to be consulted.

Incorporating UIs into administrative datasets was identified from the survey as the most frequently reported requirement to facilitate future linkage projects. However, it is important to acknowledge that evidence from the literature demonstrates that successful linkage is possible without UIs and that even using them will not ensure completely correct matching as inaccuracies are still possible through recording, transcribing and linkage protocol errors.

The introduction of EHRs in Ireland, along with the digitalising of other administrative records, has the potential to increase administrative data linkage through improved matching and linkage processes. Digital records can also be used to address some of the issues identified with administrative data quality as they can enforce standard coding systems and minimum datasets. Positive moves towards adopting these standards can be seen in the recent purchase by the HSE of a SNOMED clinical health terminology licence which had been recommend by HIQA (HIQA, 2014c). New standards and legislation will ultimately improve the overall quality of both the survey and administrative data sources and help to address the quality issues addressed in Sections 2.4.4 and 2.7.4.

While the relevant legislation, standards and ethics are all separate entities, the upcoming changes present a convergence of the principles of each, and suggest that compliance with all can be achieved with fewer resources due to the crossover. For example, conducting a PIA or equivalent is required by the GDPR, the HIQA data management standards and the new OECD ethical guidelines and, therefore completing a single PIA will assist in compliance procedures for all three.

The imminent GDPR in particular, represents a significant change in data protection legislation, and while there are several exemptions allowed for research processes within the new legislation, researchers must understand and ensure compliance in order to benefit from these exceptions. This is significant as the legislation and regulations can facilitate research by ensuring data linkage is performed correctly, but researchers must have the necessary skills and knowledge to implement the appropriate safeguards as set out in the legislation.

However, this research shows that the baseline knowledge of researchers in relation to legislation and regulation is low, which is concerning given the importance of these issues to ensure appropriate linkage procedures. There is a pronounced need to increase awareness of these issues among the research community.

The successful implementation of administrative data linkage in other countries demonstrates that it should be possible in Ireland. The findings of this research demonstrate there is significant demand from longitudinal researchers to incorporate administrative data into their projects. However, the current data linkage environment in Ireland is fragmented, complex and inconsistent, with decisions on what constitutes appropriate linkage left to individual studies and data custodians. As a result, the secondary use of administrative data is underutilised.

The evidence from this research supports the establishment of the DASSL model as an effective national infrastructure to facilitate data linkage in Ireland. This would remove the issues associated with depending on individual research projects or custodians and allow any linkage to be overseen by a national governance system. Also, establishing a national system would reduce the pressure on local resources to conduct individual data linkage and allow linkage expertise and matching technologies to develop within DASSL.

While the evidence from this research supports the establishment of DASSL, it also recognises the model will require significant time and resources to develop appropriate legislation and establish the infrastructure, and therefore there is a need to effectively support data linkage for individual studies in the interim. Failure to institute intervening support systems will further delay research and ensure that Ireland continues to lag behind international counterparts in relation to this type of research.

It is vital that during any attempted data linkage, the potential risks to privacy are addressed and minimised or mitigated where possible to ensure the societal benefits are maximised while the rights of any data subjects included in the research are upheld. However, it is important to acknowledge that if administrate data linkage is to be utilised for research purposes, ultimately the privacy risk to individuals cannot be completely eliminated. Any approaches to facilitate data linkage in Ireland must effectively balance the competing ends of research potential and privacy risk and ensure proportionate and adaptable processes and polices.

7.3. Limitations of the Research

Although this project demonstrated new evidence in relation to the benefits and challenges of data linkage in the Irish research context, there are some recognised limitations to the presented work.

7.3.1. Limitations of study methodology

Certain shortcomings in the methodology should be taken into consideration when interpreting the results of this research. The literature search was conducted on limited number of databases and therefore, there is a possibility that relevant studies, indexed in other databases, may have been missed. Additionally, this review was restricted to include only English language studies which resulted in two studies, deemed to be relevant on review of the titles/abstracts, being excluded without the full text being reviewed. Also, the inclusion criteria were applied by only one person and, to reduce the risk of bias, ideally the study selection should have been conducted independently by at least two reviewers.

While researchers from the four established longitudinal studies were the selected sample, it was not possible to rule out responses due to the 'snowballing' effect. Snowballing is a branch of convenience sampling which involves selected respondents suggesting further possible respondents to be included in the study (Weisberg, 2005). In the case of an online survey such as this, it would involve selected respondents forwarding on invite emails to other individuals or groups. In addition to affecting the set sampling frame, invite emails may be forwarded to people who do not have the required characteristics of the population of interest and can introduce error. This was highlighted as an issue particularly associated with online surveys (Van Selm and Jankowski, 2006). Due to these issues, it was not possible to calculate a response rate for the survey. This difficulty with calculating accurate response rates was highlighted by Van Selm and Jankowski (2006) as one of the main issues with online surveys.

Additionally, there are many researchers who analyse longitudinal data through public data archives such as the Irish Social Sciences Data Archive (ISSDA) which are excluded from the sampling frame of this research as they are not directly associated with a longitudinal study. This may have introduced selection bias into the research.

Researchers who completed the survey were given limited information on data linkage or what constituted administrative data before being asked questions on benefits and challenges. This may have led to misunderstanding and misinterpretation of the questions and may have affected the responses. As the questionnaire was self-administered there was limited opportunity for respondents to query any topics, although contact details of this researcher were provided these were not used by any of the respondents. For example, the responses to Question 11 indicate there may have been ambiguity in the wording as two respondents specified that they did not understand the question. However, this is a recognised weakness of self-administered questionnaires (Choi and Pak, 2005).

A sample PIA was utilised to assess how a PIA would influence a linkage project but as highlighted in Section 5.3, the results of the sample PIA were limited to the risks of the specific HIPE example and therefore the results cannot be generalised to all linkage projects – separate PIAs would be required to assess the risks of each individual potential administrative linkage project.

7.3.2. Limitations of overall study approach

Longitudinal studies were selected to review the research topic as they have been used in data linkage projects in other countries. However, the results may not be generalisable to all research studies as longitudinal studies benefit from repeated contact with participants providing an opportunity for additional consent collection if necessary (Lessof, 2009). Additionally, due to repeated contact with the research, participants in a longitudinal study are likely to have developed a level of trust with, and understanding of the project which may not be replicated in other research participants or the wider general population (Audrey et al., 2016a). Researchers working on existing cross-sectional studies aiming to link with administrative data or attempting a new study based entirely on linked administrative data would have greater difficulty contacting participants to updated consent meaning the project may not be executed if other legal methods for data processing are not possible. That is why this study set out as a review of longitudinal research only.

Researcher bias may also have potentially been introduced due to the researcher's personal perception of data linkage. The researcher works in the area of longitudinal research and has experienced the potential benefits of incorporating administrative data into longitudinal research which may have influenced the direction and conclusions of this research. However, the potential for this researcher bias is more likely to affect qualitative research and the use of mixed methods for the primary research aspect of this project should have limited the influence of this bias (Shuttleworth, 2009).

7.4. Implications for Practice and Future Research

While this study focused on the benefits and challenges of data linkage in longitudinal research projects, it is foreseen that the conclusions could be applied to enriching a wider range of research and other secondary data uses such as public health and audit and evaluation of services. As a result, the evidence identified as part of this project has provided guidance for the direction of future research and practice to enable administrative data linkage in Ireland.

While this study reviewed the opinions of longitudinal researchers, it did not include a corresponding survey of the administrative data custodians on their interpretation of the benefits and challenges of data linkage. As cooperation of both groups is needed to successfully implement data linkage, an equivalent review of the benefits, challenges and facilitators of data linkage should be conducted with administrative data custodians, particularly since custodians' unwillingness to share data was identified as a leading barrier to data linkage in the researcher survey.

Additionally, in order to include the perspectives of all stakeholders, the views and opinions of study participants and the wider general public should be reviewed. While this type of review has been conducted in other counties (Audrey et al., 2016a), there is very limited evidence of the attitudes and willingness of the Irish population to allow their personal administrative records to be incorporated into research studies, particularly if linkage is conducted in the absence of specific consent. While some information could be gleaned from the characteristic of those participants who have consented to data linkage in existing studies, a qualitative assessment of these issues is recommended in line with other international studies (Balarajan et al., 2012, Davidson et al., 2012, The Welcome Trust, 2013, Xafis, 2015). This review of public attitudes could also be utilised to collect information on the acceptance of the wider public to having their data added to a linkage system such as the DASSL model. Public engagement was identified as a key component of this model and therefore a review of public attitudes is necessary to successfully implement the proposed model.

Also, despite consent being repeatedly identified as the key method for legitimising data linkage, the review of the literature identified that the factors which influence consent to data linkage are not fully understood. This is particularly true for the Irish context as all the identified literature related to international rather than national studies. Therefore, further research is this area is required to continue to maximise participant consent and reduce and effectively control for consent bias.

In order to address the issues identified in relation to the quality of administrative datasets, data quality studies are recommended, particularly for the frequently requested datasets such as HIPE, to assess the quality and completeness of existing data and identify gaps in the current infrastructure. As well as benefiting research, this would assist administrative data custodians to identify data quality issues within their own data. While this opportunity for improving data quality may act as an incentive for administrative data custodians to engage with data linkage, it may also guide the development of the EHR programme in relation to issues of existing health data structure which could potential be addressed during the digitalisation of records. The current development of the EHR presents a unique opportunity to assess how healthcare data is structured, though such data quality studies, and identify how it can be modified to enable reuse of the data. Similar projects, such as EHR4CR, have occurred throughout Europe and have facilitated the reuse of data for clinical research (i~HD, 2017).

In relation to implications for practice, this research supports the establishment of the DASSL model. In the absence of the full model, there is a benefit to introducing elements of the model in order to maximise the data linkage that can be undertaken within the relevant legislation, ethical guidelines and standards. For example, the establishment of a national health services ethics committee similar in function to the Privacy Advisory Committee (PAC) in Scotland or the Confidentiality Advisory Group (CAG) in England could be established to review and approve proposed data linkage projects. This ethics committee could also be given statutory powers to approve the use of healthcare data in research without consent, similar to that of CAG, when justified by the public benefit. Specific national legislation such as this is permitted within the remit of the incoming GDPR.

7.5. Conclusion

This research demonstrates that there are many benefits to administrative data linkage including data correction and enrichment as well as controlling participant burden and attrition, and reducing costs. While data linkage does raise many challenges such as obtaining consent, achieving accurate matching and custodians' reluctance to engage, the evidence presented here establishes that it is possible to achieve effective linkage projects between longitudinal studies and administrative datasets. This research also shows a clear demand from researchers for more integration of administrative data into research. In order to facilitate this linkage, this research supports the establishment of a national infrastructure, in line with the proposed DASSL model, which will standardise the policies and procedures of data linkage for research and enable further administrative data linkage in Ireland.

References

ADAMS, M. M., WILSON, H. G., CASTO, D. L., BERG, C. J., MCDERMOTT, J. M., GAUDINO, J. A. & MCCARTHY, B. J. 1997. Constructing reproductive histories by linking vital records. *Am J Epidemiol,* 145**,** 339-48.

ADRN. 2017. *About the ARDN - Background* [Online]. Available: https://adrn.ac.uk/about/background/ [Accessed 5th January 2017].

ADT 2012. The UK Administrative Data Research Network: Improving Access for Research and Policy. Economic and Social Research Council.

AL BAGHAL, T. 2016. Obtaining data linkage consent for children: factors influencing outcomes and potential biases. *International Journal of Social Research Methodology,* 19**,** 623-643.

ALMEIDA, O. P., YEAP, B. B., ALFONSO, H., HANKEY, G. J., FLICKER, L. & NORMAN, P. E. 2012. Older men who use computers have lower risk of dementia. *PLoS One,* 7**,** e44239.

AUDREY, S., BROWN, L., CAMPBELL, R., BOYD, A. & MACLEOD, J. 2016a. Young people's views about consenting to data linkage: findings from the PEARL qualitative study. *BMC Med Res Methodol,* 16**,** 34.

AUDREY, S., BROWN, L., CAMPBELL, R., BOYD, A. & MACLEOD, J. 2016b. Young people's views about the purpose and composition of research ethics committees: findings from the PEARL qualitative study. *BMC Med Ethics,* 17**,** 53.

BALARAJAN, M., D'ARDENNE, J., GRAY, M. & BLAKE, M. 2012. Welsh Health Survey: Cognitive testing of data linkage consent forms and supporting documents. *In:* NATCEN (ed.).

BANKS, J., BATTY, G. D., NAZROO, J. & STEPTOE, A. 2016. The dynamics of ageing: Evidence from the English Longitudinal Study of Ageing 2002-2015 (wave 7). *In:* ELSA (ed.). UK.

BIEMER, P. P. & LYBERG, L. E. 2003. *Introduction to survey quality,* Hoboken, NJ., Wiley-Interscience.

BOYD, J. H., RANDALL, S. M., FERRANTE, A. M., BAUER, J. K., BROWN, A. P. & SEMMENS, J. B. 2014. Technical challenges of providing record linkage services for research. *BMC Med Inform Decis Mak,* 14**,** 23.

BRETT, C. E. & DEARY, I. J. 2014. Realising health data linkage from a researcher’s perspective: following up the 6-Day Sample of the Scottish Mental Survey 1947. *2014,* 5**,** 16.

BROWNELL, M. D. & JUTTE, D. P. 2013. Administrative data linkage as a tool for child maltreatment research. *Child Abuse Negl,* 37**,** 120-4.

BRYMAN, A. 2012. *Social research methods,* Oxford, Oxford University Press.

CALDERWOOD, L. & LESSOF, C. 2009. Enhancing Longitudinal Surveys by Linking to Administrative Data *In:* LYNN, P. (ed.) *Methodology of Longitudinal Surveys.* John Wiley & Sons Inc.

CALDICOTT, F. 2013. The Information Governance Review. *In:* HEALTH, D. O. (ed.). Crown.

CANNEY, C., MCNICHOLAS, T., SCARLETT, S. & BRIGGS, R. 2016. Prevalence and Impact of Chronic Debilitating Disorders. *In:* MCGARRIGLE, C., DONOGHUE, O., SCARLETT, S. & KENNY, R. (eds.) *Health and Wellbeing: Active Ageing for Older Adults in Ireland - Evidence from the Longitudinal Study on Ageing.* Dublin: TILDA.

CARROLL, M., SUTHERLAND, G., KEMP-CASEY, A. & KINNER, S. A. 2016. Agreement between self-reported healthcare service use and administrative records in a longitudinal study of adults recently released from prison. *Health Justice,* 4**,** 11.

CARTER, K., SHAW, C., HAYWARD, M. & BLAKELY, T. 2010. Understanding the determinants of consent for linkage of administrative health data with a longitudinal survey. *Kōtuitui: New Zealand Journal of Social Sciences Online,* 5**,** 53-60.

CARTER, K. N., IMLACH-GUNASEKARA, F., MCKENZIE, S. K. & BLAKELY, T. 2012. Differential loss of participants does not necessarily cause selection bias. *Aust N Z J Public Health,* 36**,** 218-22.

CHASSANG, G. 2017. The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience,* 11**,** 709.

CHATFIELD, M. D., BRAYNE, C. E. & MATTHEWS, F. E. 2005. A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies. *J Clin Epidemiol,* 58**,** 13-9.

CHOI, B. C. K. & PAK, A. W. P. 2005. A Catalog of Biases in Questionnaires. *Prev Chronic Dis.*

CHURCHES, T., CHRISTEN, P., LIM, K. & ZHU, J. X. 2002. Preparation of name and address data for record linkage using hidden Markov models. *BMC Med Inform Decis Mak,* 2**,** 9.

CNIL. 2015. *Privacy Impact Assessment - Methodology* [Online]. Available: https://www.cnil.fr/sites/default/files/typo/document/CNIL-PIA-1-Methodology.pdf [Accessed 22nd January 2017].

CORNISH, R. P., TILLING, K., BOYD, A., DAVIES, A. & MACLEOD, J. 2015. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. *Int J Epidemiol,* 44**,** 937-45.

COUNCIL OF THE EU 2015. Council Conclusions on Research Integrity - 14853/15 RECH 296. *In:* UNION, C. O. T. E. (ed.). Brussels.

COUPER, M. P., TRAUGOTT, M. W. & LAMIAS, M. J. 2001. Web Survey Design and Administration*. *Public Opinion Quarterly,* 65**,** 230-253.

CRESWELL, J. Mixed Methods Research: Design and Procedures. 2008 University of Pretoria.

CRESWELL, J. W. 2013. *Research design : qualitative, quantitative, and mixed methods approaches,* Thousand Oaks, SAGE Publications, Inc.

CSO. 2014a. *Census Small Area Data* [Online]. Available: http://www.cso.ie/en/census/census2016reports/censussmallareadata/ [Accessed 08 June 2017].

CSO. 2014b. *Officer of Statistics* [Online]. Available: http://www.cso.ie/en/aboutus/dissemination/accesstomicrodatarulespoliciesandprocedures/officersofstatistics/ [Accessed 29th May 2017].

DAVIDSON, S., MCLEAN, C., CUNNINGHAM-BURLEY, S. & PAGLIARI, C. 2012. Public Acceptability of cross sectoral data linkage: Deliberative research findings. *In:* RESEARCH, S. G. S. (ed.).

DEPOY, E. & GITLIN, L. N. 2015. *Introduction to Research: Understanding and Applying Multiple Strategies*, Mosby.

DESAI, T., RITCHIE, F. & WELPTON, R. 2016. Five Safes: Designing data access for research. Bristol: University of the West of England.

DOHC 2001. Quality and fairness: a health system for your. Dublin: The Stationary Office.

DOHC 2009. Draft Heads of Health Information Bill. Dublin: Government of Ireland.

DPC 2007. Data Protection Guidelines on research in the Health Sector. Ireland: Office of the Data Protection Commissioner.

DPC. 2017. *A Guide for Data Controllers* [Online]. Available: https://www.dataprotection.ie/docs/A-Guide-for-Data-Controllers/696.htm [Accessed 02 April 2017].

DUNN, H. L. 1946. Record Linkage. *Am J Public Health Nations Health,* 36**,** 1412-6.

EAPEN, V., WOOLFENDEN, S., WILLIAMS, K., JALALUDIN, B., DISSANAYAKE, C., AXELSSON, E. L., MURPHY, E., EASTWOOD, J., DESCALLAR, J., BEASLEY, D., CRNCEC, R., SHORT, K., SILOVE, N., EINFELD, S. & PRIOR, M. 2014. "Are you available for the next 18 months?" - methods and aims of a longitudinal birth cohort study investigating a universal developmental surveillance program: the 'Watch Me Grow' study. *BMC Pediatr,* 14**,** 234.

EISENBACH, Z., MANOR, O., PERITZ, E. & HITE, Y. 1997. The Israel Longitudinal Mortality Study--differential mortality in Israel 1983-1992: objectives, materials, methods and preliminary results. *Isr J Med Sci,* 33**,** 794-807.

ELSA. 2015. *ELSA NHSCR & Cancer Registry Consent Form - Wave 7* [Online]. Available: http://www.elsa-project.ac.uk/uploads/elsa/docs_w6/nhscr_consent_form.pdf [Accessed 01 April 2017].

ESRC 2015. Strategic Plan - 2015. Swindon, UK: Economic and Social Research Council.

EUROPEAN COMMISSION. 2016. *Protection of personal data* [Online]. Available: http://ec.europa.eu/justice/data-protection/ [Accessed 10th November 2016].

FAHCSIA 2013. Guide to Australian Longitudinal Studies. *In:* AUSTRALIAN GOVERNMENT DEPARTMENT OF FAMILIES, H., COMMUNITY SERVICES AND INDIGENOUS AFFAIRS RESEARCH AND ANALYSIS BRANCH (ed.). Canberra: Australian Government.

FLOWERS, J. & FERGUSON, B. 2010. The future of health intelligence: challenges and opportunities. *Public Health,* 124**,** 274-7.

FORD, D. V., JONES, K. H., VERPLANCKE, J.-P., LYONS, R. A., JOHN, G., BROWN, G., BROOKS, C. J., THOMPSON, S., BODGER, O., COUCH, T. & LEAKE, K. 2009. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Services Research,* 9**,** 157.

FREDMAN, L., HAWKES, W., ZIMMERMAN, S. I., HEBEL, J. R. & MAGAZINER, J. 2001. Extending Gerontological Research Through Linking Investigators' Studies to Public-Use Datasets. *The Gerontologist,* 41**,** 15-23.

GILL, L. 2001. Methods for automatic record matching and linking and their use in National Statistics (National Statistics Methodological Series No. 25). London: Office for National Statistics.

GISSLER, M. & HAUKKA, J. 2004. Finnish health and social welfare registers in epidemiological research. *Norsk epidemiologi,* 14**,** 113-120.

GOLDING, J. & JONES, R. 2009. Sources of data for a longitudinal birth cohort. *Paediatr Perinat Epidemiol,* 23 Suppl 1**,** 51-62.

GRUSKY, D. B., SMEEDING, T. M., SNIPP, C. M., JOHNSON, D. S., MASSEY, C. & O'HARA, A. 2014. The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility. *The ANNALS of the American Academy of Political and Social Science,* 657**,** 247-264.

GUEST, G., MACQUEEN, K. & NAMEY, E. 2012. *Applied thematic analysis,* Thousand Oaks, Calif. ; London, Sage.

HAGGER-JOHNSON, G. 2015. Opportunities for longitudinal data linkage in Scotland. *Scottish Medical Journal,* 61**,** 136-145.

HAGGER-JOHNSON, G., HARRON, K., GONZALEZ-IZQUIERDO, A., CORTINA-BORJA, M., DATTANI, N., MULLER-PEBODY, B., PARSLOW, R., GILBERT, R. & GOLDSTEIN, H. 2015. Identifying Possible False Matches in Anonymized Hospital Administrative Data without Patient Identifiers. *Health Serv Res,* 50**,** 1162-78.

HALPERN-MANNERS, A., WARREN, J. R. & TORCHE, F. 2014. Panel Conditioning in the General Social Survey. *Sociological Methods & Research,* 46**,** 103-124.

HARDELID, P., DATTANI, N., CORTINA-BORJA, M. & GILBERT, R. 2014. Contribution of respiratory tract infections to child deaths: a data linkage study. *BMC Public Health,* 14**,** 1191.

HART, J., HAMILTON, E. J., MAKEPEACE, A., DAVIS, W. A., LATKOVIC, E., LIM, E. M., DYER, J. R. & DAVIS, T. M. 2015. Prevalence, risk factors and sequelae of Staphylococcus aureus carriage in diabetes: the Fremantle Diabetes Study Phase II. *J Diabetes Complications,* 29**,** 1092-7.

HIQA 2010. Guidance on Privacy Impact Assessment in Health and Social Care. Dublin: Health Information and Quality Authority.

HIQA 2012. National Standards for Safer Better Healthcare. Dublin: Health Information and Quality Authority.

HIQA 2013. Guiding Principles for National Health and Social Care Data Collections. *In:* AUTHORITY, H. I. A. Q. (ed.). Dublin.

HIQA 2014a. Catalogue of National Health and Social Care Data Collections. *In:* AUTHORITY, H. I. A. Q. (ed.). Dublin.

HIQA 2014b. Recommendations for a more integrated approach to National Health and Social Care Data Collections in Ireland. *In:* DUBLIN (ed.). Health Information and Quality Authority.

HIQA 2014c. Recommendations regarding the adoption of SNOMED Clinical Terms as the Clinical Terminology for Ireland, May 2014. Dublin: Health Information and Quality Authority.

HIQA 2016. Draft Information Management standards for national health and social care data collections. Dublin: Health Information and Quality Authority.

HIQA 2017a. Five Quality Improvement Tools for National Data Collections. *In:* AUTHORITY, H. I. A. Q. (ed.). Dublin.

HIQA 2017b. Guide to the Health Information and Quality Authority's review of information management practices in national health and social care data collections. *In:* AUTHORITY, H. I. A. Q. (ed.). Dublin.

HIQA 2017c. Information Management Standards for National Health and Social Care Data Collections. *In:* AUTHORITY, H. I. A. Q. (ed.). Dublin.

HOCKLEY, C., QUIGLEY, M. A., HUGHES, G., CALDERWOOD, L., JOSHI, H. & DAVIDSON, L. L. 2008. Linking Millennium Cohort data to birth registration and hospital episode records. *Paediatr Perinat Epidemiol,* 22**,** 99-109.

HOHWÜ, L., LYSHOL, H., GISSLER, M., JONSSON, S. H., PETZOLD, M. & OBEL, C. 2013. Web-Based Versus Traditional Paper Questionnaires: A Mixed-Mode Survey With a Nordic Perspective. *Journal of Medical Internet Research,* 15**,** e173.

HOLMAN, C. D., BASS, A. J., ROSMAN, D. L., SMITH, M. B., SEMMENS, J. B., GLASSON, E. J., BROOK, E. L., TRUTWEIN, B., ROUSE, I. L., WATSON, C. R., DE KLERK, N. H. & STANLEY, F. J. 2008. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev,* 32**,** 766-77.

HPO 2015. Hospital In-Patient Enquiry (HIPE) Data Dictionary 2015, Version 7.0. *In:* OFFICE, H. P. (ed.). Dublin: Health Service Executive.

HPO 2016. Activity in Acute Public Hospitals in Ireland: 2015 Annual Report. *In:* OFFICE, H. P. (ed.). Dublin: Health Service Executive.

HRS. 2016. *HRS - Welcome!* [Online]. Available: http://hrsparticipants.isr.umich.edu/index.php?p=main&sflag=N [Accessed 11 June 2017].

HSE. 2017a. *Electronic Health Record - Background* [Online]. Available: http://www.ehealthireland.ie/Strategic-Programmes/Electronic-Health-Record-EHR-/ [Accessed 03 May 2017].

HSE. 2017b. *Individual Health Identifier - Context and Background* [Online]. Available: http://www.ehealthireland.ie/Strategic-Programmes/IHI/ [Accessed 03 May 2017].

HSE. 2017c. *The HSE and eHealth Ireland welcome the commencement order for the Individual Health Identifier* [Online]. Available: http://www.ehealthireland.ie/News-Media/News-Archive/2017/The-HSE-and-eHealth-Ireland-welcome-the-commencement-order-for-the-Individual-Health-Identifier.html [Accessed 11 June 2017].

HURE, A. J., CHOJENTA, C. L., POWERS, J. R., BYLES, J. E. & LOXTON, D. 2015. Validity and reliability of stillbirth data using linked self-reported and administrative datasets. *J Epidemiol,* 25**,** 30-7.

HUSAIN, M. J., BROPHY, S., MACEY, S., PINDER, L. M., ATKINSON, M. D., COOKSEY, R., PHILLIPS, C. J. & SIEBERT, S. 2012. HERALD (health economics using routine anonymised linked data). *BMC Med Inform Decis Mak,* 12**,** 24.

ICO. 2014. *Conducting Privacy Impact Assessments Code of Practice* [Online]. Available: https://ico.org.uk/media/for-organisations/documents/1595/pia-code-of-practice.pdf [Accessed 30th January 2017].

INTERNATIONAL EPIDEMIOLOGICAL ASSOCIATION 2014. *A dictionary of epidemiology,* Oxford, Oxford University Press.

I~HD. 2017. *About i~HD* [Online]. Available: http://www.i-hd.eu/index.cfm/about/about-i-hd/ [Accessed 19th June 2017].

JENKINS, S. P., LYNN, P., JÄCKLE, A. & SALA, E. 2008. The Feasibility of linking household survey and administrative record data: New evidence for Britain. *International Journal of Social Research Methodology,* 11**,** 29-43.

JONES, P. & ELIAS, P. 2006. Administrative data as a research resource: a selected audit. UK: The National Data Strategy.

KATZ, M. H. 2006. *Study design and statistical analysis : a practical guide for clinicians,* Cambridge, Cambridge University Press.

KAYE, J., WHITLEY, E. A., LUND, D., MORRISON, M., TEARE, H. & MELHAM, K. 2015. Dynamic consent: a patient interface for twenty-first century research networks. *European Journal of Human Genetics***,** 2.

KELLY, J. M. 2003. *The Irish Constitution,* Dublin ; London, Butterworths.

KELMAN, C. W., BASS, A. J. & HOLMAN, C. D. 2002. Research use of linked health data--a best practice protocol. *Aust N Z J Public Health,* 26**,** 251-5.

KHO, M. E., DUFFETT, M., WILLISON, D. J., COOK, D. J. & BROUWERS, M. C. 2009. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ,* 338.

KINNEAR, H., ROSATO, M., MAIRS, A., HALL, C. & O'REILLY, D. 2011. The low uptake of breast screening in cities is a major public health issue and may be due to organisational factors: a Census-based record linkage study. *Breast,* 20**,** 460-3.

KNIES, G. & BURTON, J. 2014. Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys. *BMC Medical Research Methodology,* 14**,** 125.

KNIES, G., BURTON, J. & SALA, E. 2012. Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population. *BMC Health Serv Res,* 12**,** 52.

KORBMACHER, J. M. & SCHROEDER, M. 2013. Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer. *2013,* 7**,** 17.

KREUTER, F., MÜLLER, G. & TRAPPMANN, M. 2010. Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opinion Quarterly,* 74**,** 880-906.

KUKAFKA, R., ANCKER, J. S., CHAN, C., CHELICO, J., KHAN, S., MORTOTI, S., NATARAJAN, K., PRESLEY, K. & STEPHENS, K. 2007. Redesigning electronic health record systems to support public health. *J Biomed Inform,* 40**,** 398-409.

LAMBERT, P. 2016. *Data Protection in Ireland: Sources and Issues*, Clarus Press.

LAURIE, G. & SETHI, N. 2011. Information governance of use of health-related data in medical research in Scotland: Current practices and future scenarios. Edinburgh: SHIP.

LAURIE, G. T. 2016. *Mason and McCall Smith's law and medical ethics,* Oxford, Oxford University Press.

LEONARD, H., GLASSON, E., BEBBINGTON, A., HAMMOND, G., CROFT, D., PIKORA, T., FAIRTHORNE, J., O'DONNELL, M., O'LEARY, C., HANSEN, M., WATSON, L., FRANCIS, R. W., CARTER, K. W., MCKENZIE, A., BOWER, C., BOURKE, J. & GLAUERT, R. 2013. Chapter Eight - Application of Population-Based Linked Data to the Study of Intellectual Disability and Autism. *In:* RICHARD, C. U. (ed.) *International Review of Research in Developmental Disabilities.* Academic Press.

LESSOF, C. 2009. Ethical Issues in Longitudinal Surveys. *In:* LYNN, P. (ed.) *Methodology of Longitudinal Surveys.* UK: John Wiley &amp; Sons Inc.

LINDGREN, P., JOHNSON, J., WILLIAMS, A., YAWN, B. & PRATT, G. C. 2016. Asthma exacerbations and traffic: examining relationships using link-based traffic metrics and a comprehensive patient database. *Environ Health,* 15**,** 102.

LINSTONE, H. & TUROFF, M. 2002. *The Delphi Method: Techniques and Applications,* United States.

LYNN, P. 2009. Methods for Longitudinal Surveys. *In:* LYNN, P. (ed.) *Methodology of Longitudinal Surveys.* UK: John Wiley & Sons Inc.

MACFEELY, S. & DUNNE, J. 2014. Joining up public service information: The rationale for a national data infrastructure. *Administration,* 61**,** 93-107.

MACLEOD, J., COPELAND, L., HICKMAN, M., MCKENZIE, J., KIMBER, J., DE ANGELIS, D. & ROBERTSON, J. R. 2010. The Edinburgh Addiction Cohort: recruitment and follow-up of a primary care based sample of injection drug users and non drug-injecting controls. *BMC Public Health,* 10**,** 101.

MARS, B., CORNISH, R., HERON, J., BOYD, A., CRANE, C., HAWTON, K., LEWIS, G., TILLING, K., MACLEOD, J. & GUNNELL, D. 2016. Using Data Linkage to Investigate Inconsistent Reporting of Self-Harm and Questionnaire Non-Response. *Arch Suicide Res,* 20**,** 113-41.

MARTIN, J., BYNNER, J., KALTON, G., BOYLE, P., GOLDSTEIN, H., GAYLE, V., PARSONS, S. & PIESSE, A. 2006. Strategic Review of Panel and Cohort Studies: Report to the Research Resources Board of the Economic and Social Research Council. London: Longview.

MAXWELL, J. 1969. *Sixteen years on : a follow-up of the 1947 Scottish survey,* London, University of London P.

MCGHEE, J., MITCHELL, F., DANIEL, B. & TAYLOR, J. 2015. Taking a Long View in Child Welfare: How Can We Evaluate Intervention and Child Wellbeing Over Time? *Child Abuse Review,* 24**,** 95-106.

MCGIVERN, Y. 2006. *The practice of market and social research : an introduction,* Harlow, Financial Times Prentice Hall.

MCKINNEY, P. A., JONES, S., PARSLOW, R., DAVEY, N., DAROWSKI, M., CHAUDHRY, B., STACK, C., PARRY, G. & DRAPER, E. S. 2005. A feasibility study of signed consent for the collection of patient identifiable information for a national paediatric clinical audit database. *Bmj,* 330**,** 877-9.

MICHAUD, S., DOLSON, D., ADAMS, D. & RENAUD, M. 1995. *Combining administrative and survey data to reduce respondent burden in longitudinal surveys*, Survey of Labour and Income Dynamics, Statistics Canada.

MORAN, R. 2016. Proposals for an Enabling Data Environment for Health and Related Research in Ireland. Dublin: Health Research Board.

MORIARTY, F., BENNETT, K., FAHEY, T., KENNY, R. A. & CAHIR, C. 2015. Longitudinal prevalence of potentially inappropriate medicines and potential prescribing omissions in a cohort of community-dwelling older people. *Eur J Clin Pharmacol,* 71**,** 473-82.

MOSTAFA, T. 2016. Variation within households in consent to link survey data to administrative records: evidence from the UK Millennium Cohort Study. *International Journal of Social Research Methodology,* 19**,** 355-375.

MOUNTAIN, J. A., NYARADI, A., ODDY, W. H., GLAUERT, R. A., DE KLERK, N. H., STRAKER, L. M. & STANLEY, F. J. 2016. Data linkage in an established longitudinal cohort: the Western Australian Pregnancy Cohort (Raine) Study. *Public Health Res Pract,* 26.

NSB 2011. Joined up Government Needs Joined Up Data. Dublin: Government of Ireland Stationary Office.

NSB 2015. A World Class Statistics System for Ireland: Strategic Priorities for Official Statistics 2015-2020. Dublin: Government of Ireland.

OECD 2016. Research Ethics and New Forms of Data for Social and Economic Research *OECD Science, Technology and Industry Policy Papers* Paris: OECD Publishing.

PAOLO, A. M., BONAMINIO, G. A., GIBSON, C., PARTRIDGE, T. & KALLAIL, K. 2000. Response rate comparisons of e-mail- and mail-distributed student evaluations. *Teach Learn Med,* 12**,** 81-4.

PONTO, J. 2015. Understanding and Evaluating Survey Research. *J Adv Pract Oncol,* 6**,** 168-71.

PUDNEY, S. 2008. Heaping and Leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure. UK: ISER.

RAJULTON, F. 2001. The Fundamentals of Longitudinal Research: An Overview. . *Canadian Studies in Population,* 28**,** 169-185.

RENZI, C., LYRATZOPOULOS, G., CARD, T., CHU, T. P., MACLEOD, U. & RACHET, B. 2016. Do colorectal cancer patients diagnosed as an emergency differ from non-emergency patients in their consultation patterns and symptoms? A longitudinal data-linkage study in England. *Br J Cancer,* 115**,** 866-75.

ROBIN, A. 1992. *Learning about, diagnosing and communicating error in longitudinal panel surveys,* [Colchester], University of Essex.

SAFRAN, C., BLOOMROSEN, M., HAMMOND, W. E., LABKOFF, S., MARKEL-FOX, S., TANG, P. C., DETMER, D. E. & EXPERT, P. 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**,** 1-9.

SAKSHAUG, J., TUTZ, V. & KREUTER, F. 2013. Placement, Wording, and Interviewers: Identifying Correlates of Consent to Link Survey and Administrative Data. *Survey Research Methods; Vol 7, No 2 (2013)*.

SAKSHAUG, J. W. & KREUTER, F. 2012. Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data. *2012,* 6**,** 10.

SALA, E., BURTON, J. & KNIES, G. 2012. Correlates of Obtaining Informed Consent to Data Linkage: Respondent, Interview, and Interviewer Characteristics. *Sociological Methods & Research,* 41**,** 414-439.

SALA, E., KNIES, G. & BURTON, J. 2014. Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology,* 17**,** 455-473.

SALMAN, R. A.-S., BELLER, E., KAGAN, J., HEMMINKI, E., PHILLIPS, R. S., SAVULESCU, J., MACLEOD, M., WISELY, J. & CHALMERS, I. 2014. Increasing value and reducing waste in biomedical research regulation and management. *The Lancet,* 383**,** 176-185.

SARIYAR, M. & SCHLUNDER, I. 2016. Reconsidering Anonymization-Related Concepts and the Term "Identification" Against the Backdrop of the European Legal Framework. *Biopreserv Biobank,* 14**,** 367-374.

SCHUTZE, R. 2015. *An introduction to European law,* Cambridge, United Kingdom, Cambridge University Press,.

SEEMATTER-BAGNOUD, L. & SANTOS-EGGIMANN, B. 2006. Population-based cohorts of the 50s and over: a summary of worldwide previous and ongoing studies for research on health in ageing. *European Journal of Ageing,* 3**,** 41.

SHEEHAN, K. B. 2001. E-mail Survey Response Rates: A Review. *Journal of Computer-Mediated Communication,* 6**,** 0-0.

SHUTTLEWORTH, M. 2009. *Researcher Bias* [Online]. Available: https://explorable.com/research-bias [Accessed 01 June 2017].

SILLS, S. J. & SONG, C. 2002. Innovations in Survey Research. *Social Science Computer Review,* 20**,** 22-30.

SOLOFF, C., SANSON, A., WAKE, M. & HARRISON, L. 2007. Enhancing Longitudinal Studies by Linkage to National Databases: Growing Up in Australia, the Longitudinal Study of Australian Children. *International Journal of Social Research Methodology,* 10**,** 349-363.

STONE, M. A., REDSELL, S. A., LING, J. T. & HAY, A. D. 2005. Sharing patient data: competing demands of privacy, trust and research in primary care. *Br J Gen Pract,* 55**,** 783-9.

SULLIVAN, F., MCKINSTRY, B. & PALMER, C. 2016. Opt-in method is vital for data sharing. *Bmj,* 354**,** i4293.

TATE, A. R., CALDERWOOD, L., DEZATEUX, C. & JOSHI, H. 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**,** 294-298.

THE WELCOME TRUST 2013. Summary Report of Qualitative Research into Public Attitudes to Personal Data and Linking Personal Data. London.

THORNTON, I. & HIPSKIND, J. 2017. Informed Consent. *StatPearls.* Treasure Island (FL): StatPearls PublishingStatPearls Publishing LLC.

THORNTON, M., WILLIAMS, J., MCCRORY, C., MURRAY, A. & QUAIL, A. 2013. Growing Up in Ireland: Design, Instrumentation and Procedures for the Infant Cohort at Wave One (9 months). *In:* GUI (ed.). Dublin: ESRI.

TICKNER, N. 2013. School Completers - What's Next? Report on School Completers from Post-Primary Schools - pupils enrolled in 2009/2010 and not in 2010/2011. *In:* SKILLS, D. O. E. A. (ed.). Ireland.

TILDA. 2017. *Where are we now?* [Online]. Available: http://tilda.tcd.ie/about/where-are-we-now/ [Accessed 11 June 2017].

TOOTH, L. R., HOCKEY, R., TRELOAR, S., MCCLINTOCK, C. & DOBSON, A. 2012. Does government subsidy for costs of medical and pharmaceutical services result in higher service utilization by older widowed women in Australia? *BMC Health Serv Res,* 12**,** 179.

TOURANGEAU, R., RIPS, L. R. & RASINSKI, K. 2000. *The psychology of survey response,* Cambridge, U.K., Cambridge University Press.

TOWNSLEY, R. 2016. ESRC Longitudinal Studies Review 2017 Interim report: Initial analysis of responses to the consultation. *In:* ESRC (ed.). Swindon, UK.

UKDF 2013. UK Strategy for Data Resources for Social and Economic Research 2013-2018. Swindon, UK: UK Data Forum.

VAN SELM, M. & JANKOWSKI, N. W. 2006. Conducting Online Surveys. *Quality and Quantity,* 40**,** 435-456.

WALLEY, T. 2006. Using personal health information in medical research: Overzealous interpretation of UK laws is stifling epidemiological research. *Bmj.*

WATSON, N. & WOODEN, M. 2009. Identifying Factors Affecting Longitudinal Survey Response. *In:* LYNN, P. (ed.) *Methodology of Longitudinal Surveys.* UK: John Wiley & Sons Inc.

WEISBERG, H. F. 2005. *The total survey error approach : a guide to the new science of survey research,* Chicago, Ill. ; London, University of Chicago Press,.

WHELAN, B. J. & SAVVA, G. M. 2013. Design and methodology of the Irish Longitudinal Study on Ageing. *J Am Geriatr Soc,* 61 Suppl 2**,** S265-8.

WILLISON, D. 2009. Use of Data from the Electronic Health Record for Health Research - current governance challenges and potential approaches. Ontario, Canada.

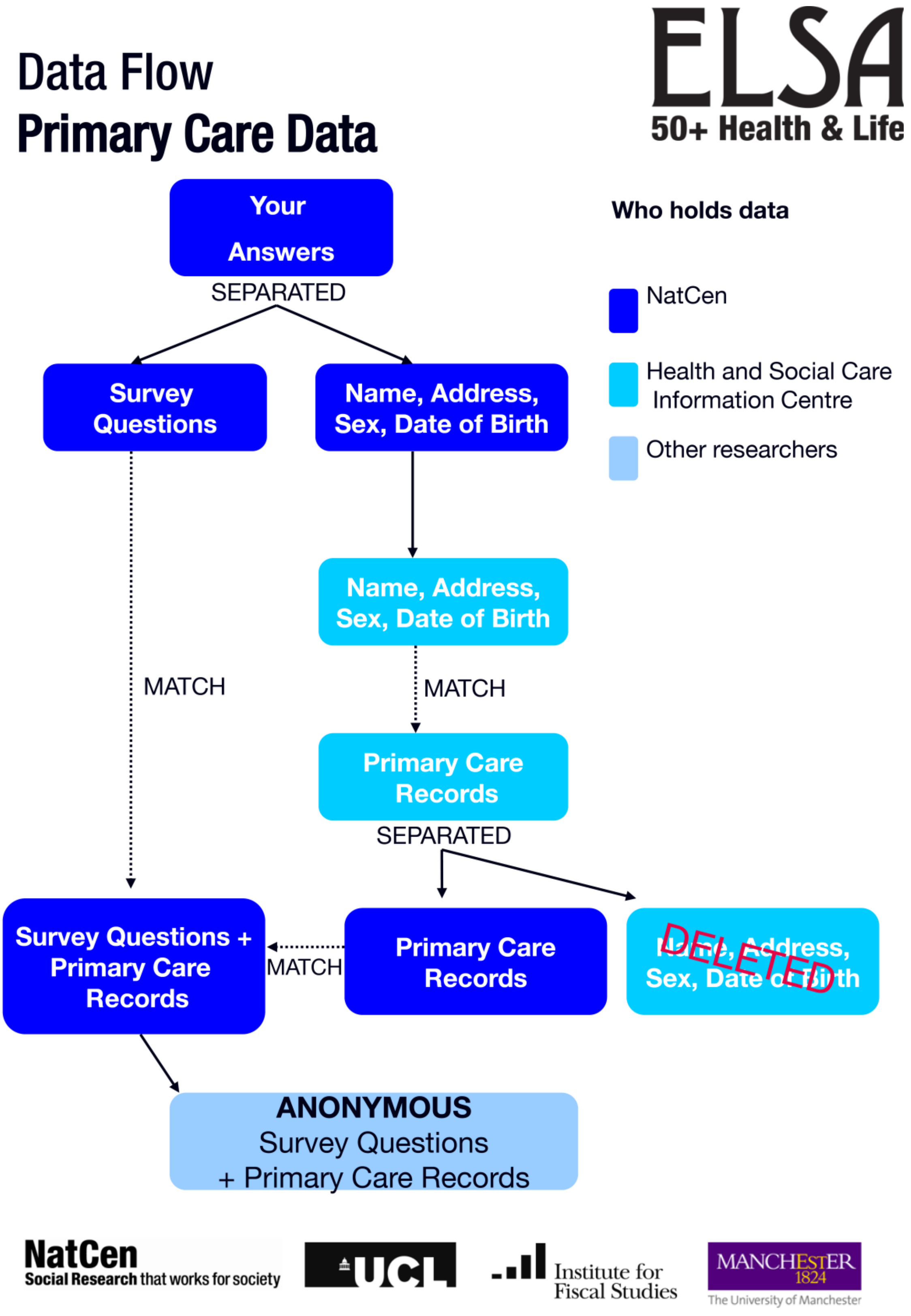
XAFIS, V. 2015. The acceptability of conducting data linkage research without obtaining consent: lay people’s views and justifications. *BMC Medical Ethics,* 16**,** 79.

YAN, T. & CURTIN, R. 2010. The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research,* 22**,** 535-551.

YEAP, B. B., ALFONSO, H., HANKEY, G. J., FLICKER, L., GOLLEDGE, J., NORMAN, P. E. & CHUBB, S. A. 2013. Higher free thyroxine levels are associated with all-cause mortality in euthyroid older men: the Health In Men Study. *Eur J Endocrinol,* 169**,** 401-8.

YIN, R. K. 2009. *Case Study Research: Design and Methods*, SAGE Publications.

Appendix A: ELSA Data Linkage Participant Information Sheet



Data Flow
Primary Care Data

ELSA
50+ Health & Life

Who holds data

■ NatCen

■ Health and Social Care Information Centre

■ Other researchers

Your Answers

SEPARATED

Survey Questions

Name, Address, Sex, Date of Birth

MATCH

Name, Address, Sex, Date of Birth

MATCH

Primary Care Records

SEPARATED

Survey Questions + Primary Care Records

MATCH

Primary Care Records

Name, Address, Sex, Date of Birth

DELETED

ANONYMOUS
Survey Questions
+ Primary Care Records

NatCen Social Research that works for society

UCL

Institute for Fiscal Studies

MANCHESTER 1824
The University of Manchester

Appendix B: Sample Privacy Impact Assessment: Data Linkage of Longitudinal Study Data with HIPE Data

Introduction

The Sample Longitudinal Study (SLS) is a birth cohort longitudinal study which follows all aspects of the health, economic and social circumstances of over 8,000 individuals born in 1950 in Ireland. Participants are contacted every three years to complete an in-depth interview followed by a comprehensive health assessment.

The Hospital In-Patient Enquiry (HIPE), which is managed by the Healthcare Pricing Office (HPO), collects clinical and administrative data on all discharges from and deaths in acute hospitals in Ireland (HPO, 2015). The HIPE database contains information on both day patient and in-patient activity in the acute public hospitals (HPO, 2016).

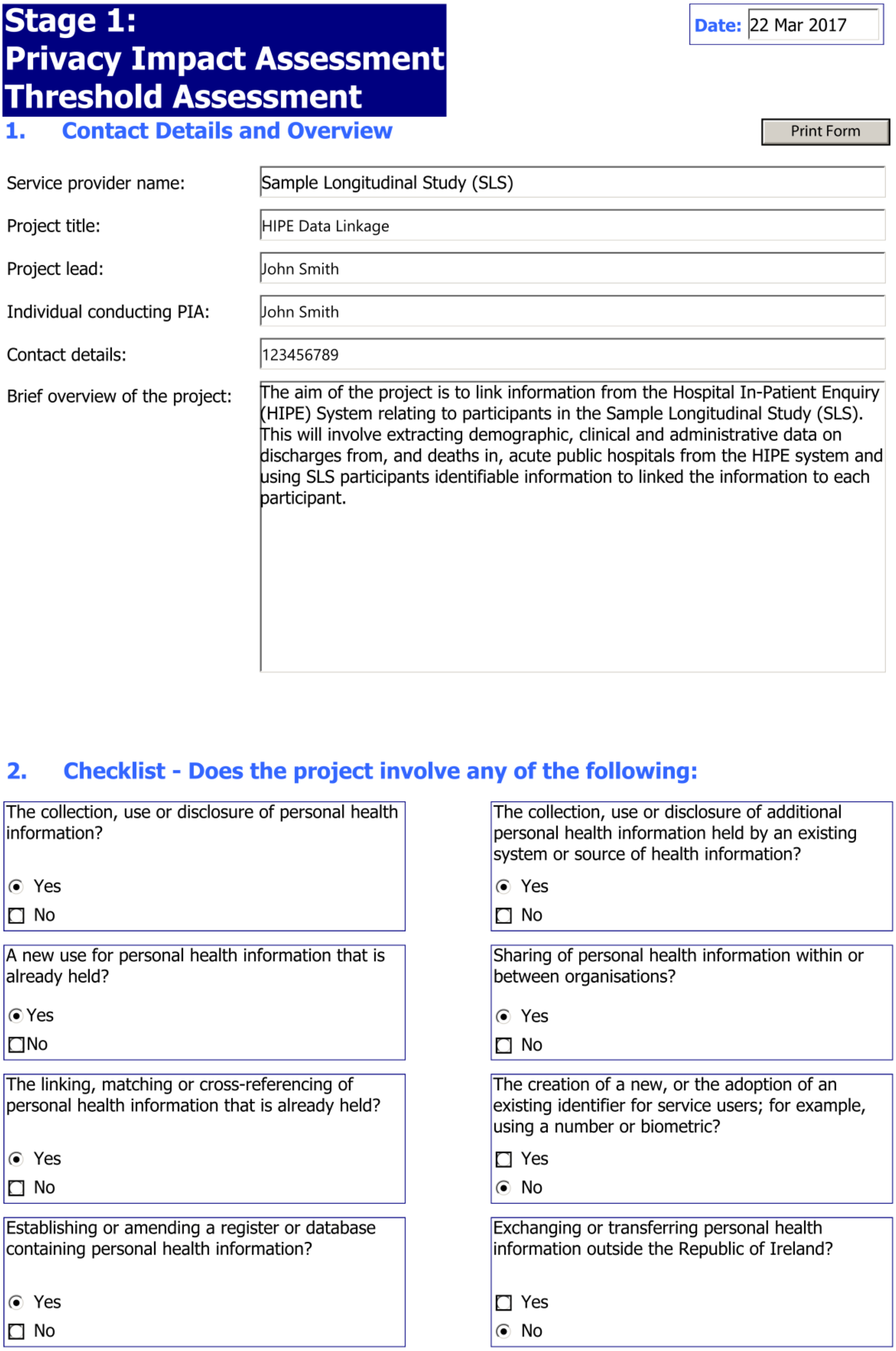
In an effort to enrich the data currently collected in SLS, a linkage project has been proposed with the HIPE system. This project would involve supplementing the SLS data with clinical information and healthcare utilisation data for individual participants from the HIPE system. The linkage will be dependent on informed written consent from each participant which will be collected during the next round of SLS interviews. Any participants who do not consent to data linkage will be excluded from the project.

This PIA will not explore the privacy risks associated with either the HIPE system or the SLS individually, instead focusing on the privacy risks associated with the sharing and linkage of data between the two.

This PIA utilised the HIQA guidance document on conducting PIAs and the stages and review questions outlined below reflect the recommended process as outlined by HIQA (HIQA, 2010).

Stage 1 – PIA threshold assessment

As recommended by HIQA (HIQA, 2010), the first stage of the PIA process was to conduct a threshold assessment to assess the requirement for a PIA. Based on this threshold assessment (Box X), it was determined that the linkage project had the potential to impact participants' privacy and a full PIA was necessitated.



Stage 1: Privacy Impact Assessment Threshold Assessment

Date: 22 Mar 2017

1. Contact Details and Overview

Print Form

Service provider name:	Sample Longitudinal Study (SLS)
Project title:	HIPE Data Linkage
Project lead:	John Smith
Individual conducting PIA:	John Smith
Contact details:	123456789
Brief overview of the project:	The aim of the project is to link information from the Hospital In-Patient Enquiry (HIPE) System relating to participants in the Sample Longitudinal Study (SLS). This will involve extracting demographic, clinical and administrative data on discharges from, and deaths in, acute public hospitals from the HIPE system and using SLS participants identifiable information to linked the information to each participant.

2. Checklist - Does the project involve any of the following:

The collection, use or disclosure of personal health information?

◉ Yes
□ No

The collection, use or disclosure of additional personal health information held by an existing system or source of health information?

◉ Yes
□ No

A new use for personal health information that is already held?

◉ Yes
□ No

Sharing of personal health information within or between organisations?

◉ Yes
□ No

The linking, matching or cross-referencing of personal health information that is already held?

◉ Yes
□ No

The creation of a new, or the adoption of an existing identifier for service users; for example, using a number or biometric?

□ Yes
◉ No

Establishing or amending a register or database containing personal health information?

◉ Yes
□ No

Exchanging or transferring personal health information outside the Republic of Ireland?

□ Yes
◉ No

The use of personal data for research or statistics, whether de-identified or not?

◉ Yes
□ No

A new or changed system of data handling; for example, policies or practices around access, security, disclosure or retention of personal health information?

◉ Yes
□ No

Any other measures that may affect privacy or that could raise privacy concerns with the public?

◉ Yes
□ No

If the answer to one or more of the questions is "yes" then a Privacy Impact Assessment must be undertaken. If the answer to all of the questions is "no" it will not be necessary to complete a Privacy Impact Assessment.

3. Recommendation

Individual conducting the threshold assessment:

A Privacy Impact Assessment:

◉ is required
□ is not required

Name: John Smith
Signature: John Smith
Title: SLS Research Manager
Date: 22 Mar 2017

Endorsement by senior management:

Privacy Impact Assessment recommendation:

◉ Agree
□ Disagree

Name: Jane Doe
Signature: Jane Doe
Title: SLS Principle Investigator
Date: 22 Mar 2017

Stage 2 – Identification of privacy risk

Privacy Management: Table B.1 outlines the status of practices and policies of the SLS relating to privacy management. The SLS currently has a compressive data management policy in place. This policy address issues such as approved uses of collected data, staff obligations in relation to data management and protection, data security, data retention and destruction and data breach management. All staff are required to undergo training and sign the data management policy to indicate understanding and acceptance before accessing any of the study data. There is also a set governance structure in place with clearly defined roles and responsibilities in relation to data management and protection.

Table B.1: Privacy management in the SLS

Privacy Management Issue	Status
Is there a privacy policy in place?	Yes
Is there a statement of information practices?	Yes
Is the study compliant with data protection legislation?	Yes
Is the study the legal data controller for all personal data currently being processed?	Yes
Is there a records management policy in place that includes a retention and destruction schedule?	Yes
Are administrative, technical and physical safeguards in place to protect personal information against theft, loss, unauthorised use or disclosure and unauthorised copying, modification or disposal?	Yes
Is there an appointed privacy or information governance contact person?	Yes
Is there a privacy breach management action plan in place?	Yes
Are employees or agents with access to personal health information provided with training related to privacy protection and confidentiality requirements?	Yes

Project Description: As detailed in the PIA introduction, this linkage project will involve identifying SLS participants who are included in in the HIPE database, returning HIPE data relating to participants to SLS and then linking this HIPE data with existing information collected as part of the research study. The aim of the data linkage is to enrich the existing SLS data with clinical data such as diagnoses and treatments as well as information relating to healthcare utilisation.

The sharing and linkage of HIPE data will be dependent on written informed consent which will be collected at the next wave of SLS interviews. Once a participant has consented to data linkage, their name, address, sex and DOB along with a unique linkage identifier will be securely sent to HPO as the HIPE data custodians. This information will be used to identify SLS participants in the HIPE database and relevant clinical and utilisation information will be extracted. This information will be securely returned to SLS with the unique linkage identifier included. Name, address, sex and DOB will all be removed from the data by HPO in advance of returning information to SLS. A summary of the data to be shared and linked as part of the project is detailed in Table B.2.

Table B.2: Proposed data for inclusion in linkage project

Data shared from SLS to HIPE	Participant name Participant address Participant sex Participant date of birth
Data shared from HIPE to SLS	Patient status (public, private) Admission date Admission type (emergency, elective, maternity, etc.) Source of admission (home, nursing home, other hospital, etc.) Discharge date Discharge status (transfer to home, nursing home, other hospital or death) Consultant speciality Diagnosis Treatment and procedure details Intensive care treatment details

Scope of the project: Table B.3 presents an overview of the scope of the proposed linkage project along with any potential privacy risks identified. The aim of this stage of the PIA is to review the proposed uses of personal data and why its use is necessary. A series of questions recommended by HIQA are used to review the scope of the project, the answers of which reflect the related processes and any safeguards which will be in place. Table B.3 also summarises any potential privacy risks associated with the processes.

Table B.3: Scope of the proposed project and associated privacy risks

PIA Question	Answer	Privacy Risk
What information will be collected in the proposed project?	No additional information will be collected as part of this linkage project. The project will instead involve bidirectional sharing of data between SLS and HIPE. This will result in HIPE being made aware of SLS participation and SLS having access to patient healthcare information.	Both organisations will have personal data disclosed to them, of which they are not the data controllers. There is a risk that information will be used inappropriately or disclosed further to additional parties. There is also a risk that staff in either SLS or HIPE may know the individual to which the data refers. However, staff in both organisations are bound by contractual confidentiality clauses.
Outline why each element of the dataset is necessary	Identifiable information from SLS is required by HIPE to effective identify participants in the HIPE dataset. The lack of a UHI necessitates the sharing of this information. The information shared by HIPE with SLS, as outlined in Table B.2), is required to effectively address several research questions and also to ensure sufficient data to fully address any future research questions or theories.	Not providing sufficient data from SLS to HIPE may result in inefficient matching and requesting insufficient data from HIPE to SLS may limit the research that can be conducted on the linked data. Both of these scenarios may require the linkage process be repeated which increases the opportunity for error or privacy risk.

PIA Question	Answer	Privacy Risk
Are the data subjects aware of the proposed collection, use and disclosure of their personal information? Identify and describe what information is given and how it is given.	Yes - all participants will be informed of the linkage project through an information booklet and informed consent form. SLS interviewers will be fully trained in relation to the linkage project meaning participants have the opportunity to ask questions about the project during the consent process.	Risk that information booklet and consent form will not sufficiently notify participants about how their data will be used meaning consent will not be fully informed. Requirement that information booklet and consent form are compliant with data protection legislation.
Have the data subjects consented to their personal information being used in this manner? Describe the consent process.	Yes – informed written consent will be collected during the SLS interview. Participants will receive an information booklet and a copy of the signed consent form. No data linkage will be performed for participants who do not consent.	Process required to ensure information of participants who do not consent are not inadvertently sent to HIPE.
Identify and describe: • All the uses of the personal information • How these uses relate to the purpose for which the information was collected • Any changes to the purpose for using the information after the information is collected • Measures in place to prevent use for secondary purposes	The linked data will be used solely for research purposes by SLS. Use of the data will not be limited to a single research question but use will be limited to the SLS research team and all analysis will be reviewed and approved by SLS management. This use of the HIPE data is beyond the purpose for which it was originally collected and therefore constitutes a secondary use of the data.	This constitutes a new use for data beyond that for which it was originally collected. This reuse of data must be legally compliant.

PIA Question	Answer	Privacy Risk
	The use of the linked data will be restricted to SLS and will not be disclosed further beyond the SLS team.	
Identify and describe any potential sharing of the information and how the data subject has been informed of this.	The project revolves around the sharing of data between SLS and HIPE. Data subjects will be informed through an information booklet and consent from.	Wording of consent may not be sufficient to ensure HIPE can legally disclose personal health information to SLS. Consent wording must account for any confidentiality assurances that were made by HIPE to patients at the point of collection.
Is it a possibility that the information will be linked or matched with an existing or proposed system? If yes please provide details	Data will be linked as this is the main aim of the project. Participants will be informed that their HIPE data will be linked to the information that they have previously provided to SLS.	Risk of incorrect or inaccurate linkage occurring between the SLS and HIPE datasets. This would result in incorrect information being associated with a participant.
Does the project, system or initiative involve assigning or using an identifier or using an existing identifier for a new purpose?	The project will involve the generation of a linkage identifier. This will be generated and provided by SLS and retuned by HIPE with the accompanying health record data. The use of this identifier means returned HIPE data will not contain names, address, etc. but the information can still be linked back to the SLS study data. This identifier will not be used for any other purpose	No additional risks identified.

Information flows: The flow of personal information for the linkage process is outlined in Figure B.1. Privacy relevant issues and associated risks in relation to the information flow are discussed in Table B.4.

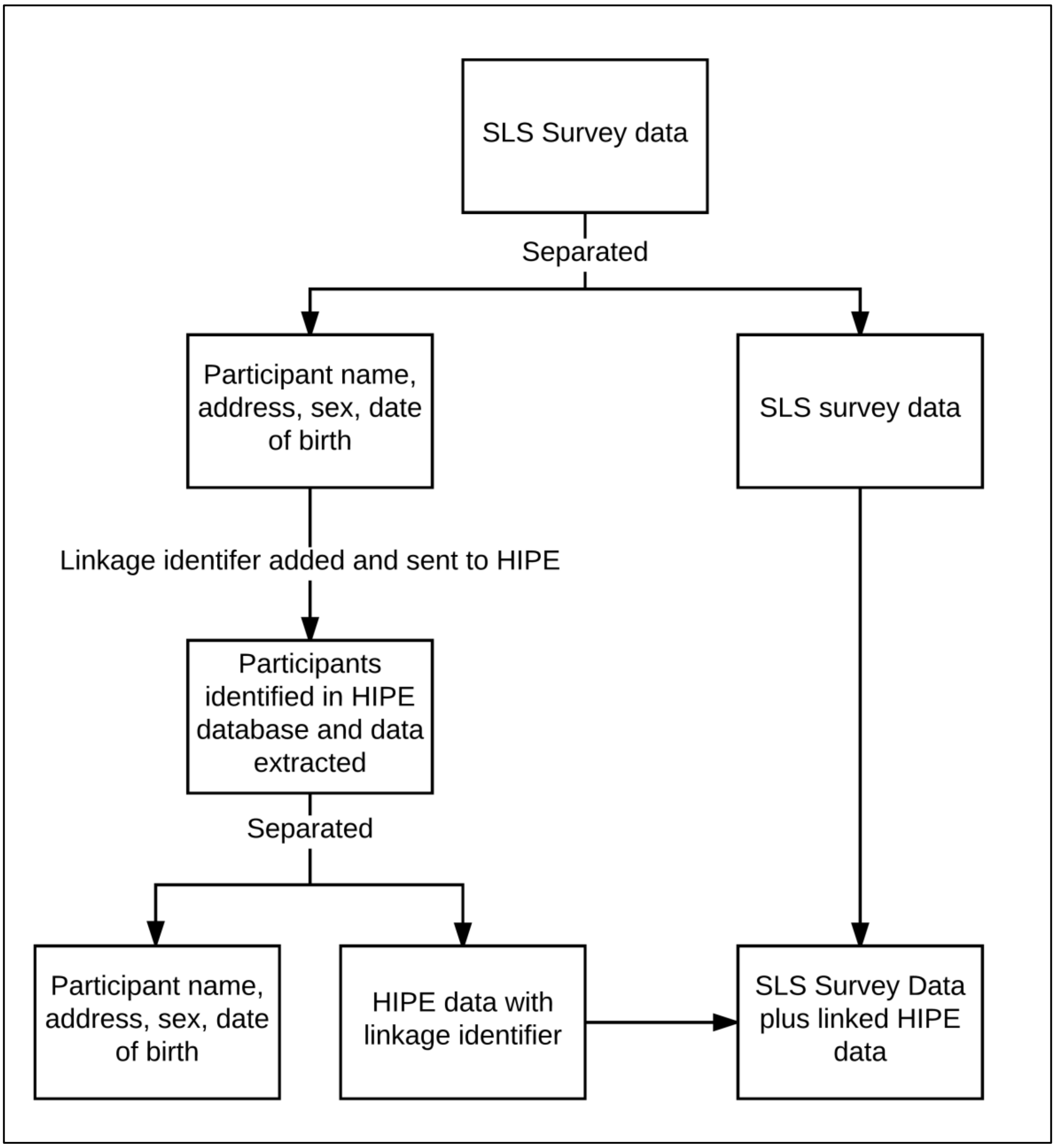

SLS Survey data
Separated
Participant name, address, sex, date of birth
SLS survey data
Linkage identifer added and sent to HIPE
Participants identified in HIPE database and data extracted
Separated
Participant name, address, sex, date of birth
HIPE data with linkage identifier
SLS Survey Data plus linked HIPE data

Figure B.1: Proposed data linkage information flow

Table B.4: Proposed information flow and associated privacy risks

PIA Question	Answer	Privacy Risk
How will the information be collected?	No new information will be collected. The project will involve sharing and linkage of data between SLS and HIPE.	No additional risks identified.
What are the proposed uses of the information?	The data will be used solely for academic and policy research. Information will only be published in aggregated forms and it will not be possible to identify any participants from published results.	No additional risks identified.
Will the information be disclosed? To who? What precautions are in place?	Information collected by both SLS and HIPE will be shared with the other organisation. Staff in both organisations are bound by contractual confidentiality clauses. Secure encrypted portals will be used for all data transfers. Minimum data storage security will be agreed in advance of any data transfer. Identifiable information provided by SLS to HIPE will be deleted once health record data has been extracted and transferred to SLS.	Information being disclosed to third party by either SLS or HIPE.

PIA Question	Answer	Privacy Risk
Will the data subjects have access to the information and have the opportunity to have any information about them corrected?	Both SLS and HIPE are required to comply with subject access requests. The participant name, address, sex and date of birth data provided by SLS will be deleted by HIPE once health record data has been extracted and transferred to SLS and therefore there will be no data for data subjects to request access or correction.	Risks associated with responsibility in relation to data access requests. SLS will need to disclose the HIPE data to the participant which it relates to if a subject access request is received. Agreement also needed for how to process a request from data subjects to correct HIPE data held by SLS.
What security measures will be taken to protect the information from loss, unauthorised access, use, modification, disclosure or other misuse, including how data is transferred from sites?	The HIPE data will be incorporated into the existing SLS dataset and will be subject to the same rigorous data management policies and procedures. These set standards for how data should be accessed and use as well as detailing back-up and disaster recovery procedures to prevent data loss. Participant identifiable information is stored separately to survey and other research data with only key senior staff having access to the linkage key. Research data is also stored in a read-only format to prevent modification of the data. Data will be transferred via secure encrypted portal	No additional risks identified.

PIA Question	Answer	Privacy Risk
Identify and describe the retention and destruction practices to be employed in the project	The participant name, address, sex and date of birth data provided by SLS will be deleted by HIPE once health record data has been extracted and transferred to SLS. Data provided by HIPE to SLS will be incorporated into existing SLS data any comply with SLS retention and destruction polies. As SLS is a longitudinal study, all data is to be retained for the lifetime of the project. Once primary data collection is completed all study data will be retained in an anonymised form to ensure continued use of the research data. Any linkage key files between SLS data and participants' identities will be deleted.	No additional risks identified.

Summary of identified risks: Based on the review carried out above, crossover of some potential risks were identified and therefore have been grouped together. The finalised risks are detailed below:

1. This project constitutes a new use for data beyond that for which it was originally collected and this reuse of data must be legally compliant. There is a risk that the information booklet and consent form will not sufficiently notify participants about how their data will be used meaning consent will not be fully informed. Additionally, the wording of the consent may not be sufficient to ensure HIPE can legally disclose personal health information to SLS.

2. Both parties will have personal data disclosed to them, of which they are not the data controllers. There is a risk that information will be used inappropriately or disclosed further to additional parties.

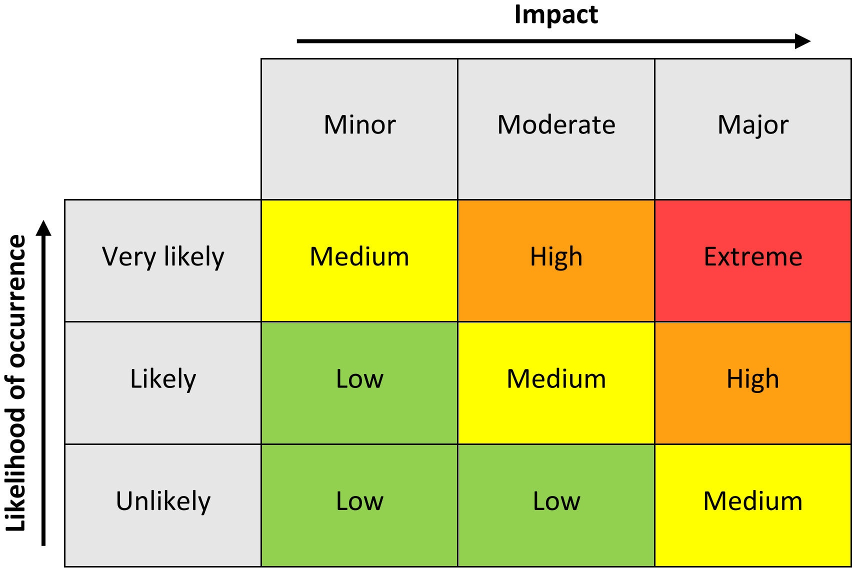
3. Not providing sufficient data from SLS to HIPE may result in inefficient matching and requesting insufficient data from HIPE to SLS may limit the research that can be conducted on the linked data. Both of these scenarios may require the linkage process be repeated which increases the opportunity for error or privacy risk.

4. If an inadequate linkage process is conducted there is a risk of incorrect matching which will result in incorrect information being assigned to participants. There is also a risk that details of participants who did not consent to inclusion in the project will be incorrectly sent to HIPE.

5. Risks associated with responsibility in relation to data access requests. SLS will need to disclose the HIPE data to the participant which it relates to if a subject access request is received.

Stage 3 – Addressing Privacy Risks

A risk matrix (Figure B.2) was used to analyse and classify the identified risks. This matrix rates each risk based on the likelihood of it occurring and the resulting impact it would have. The risk rating was used to determine the appropriate level of management or intervention that would be required to address each of the risks.


Impact
Likelihood of occurrence

	Minor	Moderate	Major
Very likely	Medium	High	Extreme
Likely	Low	Medium	High
Unlikely	Low	Low	Medium

Figure B.2: PIA risk matrix (Source: HIQA, 2010)

Each of the privacy risks identified in stage 2 were evaluated in relation to their likelihood of occurrence and they impact they would have if they were to occur. Details of this evaluation are detailed in Table B.5.

Table B.5: Privacy risk rating

ID	Risk	Likelihood	Impact	Risk Rating
1	This project constitutes a new use for data beyond that for which it was originally collected and this reuse of data must be legally compliant. There is a risk that the information booklet and consent form will not sufficiently notify participants about how their data will be used meaning consent will not be fully informed. Additionally, the wording of the consent may not be sufficient to ensure HIPE can legally disclose personal health information to SLS.	Unlikely	Major	Medium
2	Both parties will have personal data disclosed to them, of which they are not the data controllers. There is a risk that information will be used inappropriately or disclosed further to additional parties.	Unlikely	Major	Medium
3	Not providing sufficient data from SLS to HIPE may result in inefficient matching and requesting insufficient data from HIPE to SLS may limit the research that can be conducted on the linked data. Both of these scenarios may require the linkage process be repeated which increases the opportunity for error or privacy risk.	Likely	Moderate	Medium
4	If an inadequate linkage process is conducted there is a risk of incorrect matching which will result in incorrect information being assigned to participants. There is also a risk that details of participants who did not consent to inclusion in the project will be incorrectly sent to HIPE.	Likely	Moderate	Medium
5	Risks associated with responsibility in relation to data access requests. SLS will need to disclose the HIPE data to the participant which it relates to if a subject access request is received.	Unlikely	Minor	Low

Once the identified risks had been analysed and rated, each of the risks were reviewed to develop interventions to reduce or eliminate them. Each of the proposed solutions were evaluated to determine if the associated risks would be eliminated, reduced or accepted. Any efforts to address the risks were balanced against the aims of the linkage project. The introduction of new risks arising from the proposed solutions were also considered. Each of the risks and the proposed solutions to address them are detailed in Table B.6.

Table B.6: Privacy risks and proposed solutions

<table>
<tr><td colspan="2"><b>Risk 1:</b> This project constitutes a new use for data beyond that for which it was originally collected and this reuse of data must be legally compliant. There is a risk that the information booklet and consent form will not sufficiently notify participants about how their data will be used meaning consent will not be fully informed. Additionally, the wording of the consent may not be sufficient to ensure HIPE can legally disclose personal health information to SLS.</td></tr>
<tr><td><b>Proposed solution:</b> Legal compliance for the reuse of HIPE data will be achieved through participant consent. The information booklet and consent form will be reviewed and approved by both an academic ethics committee and a data protection lawyer in advance of being provided to participants. The consent process will be piloted on a minimum of 50 SLS participants to ensure wording is clear and easily understood.
A legally binding contract between SLS and HIPE will be developed and agreed in advance of seeking consent for linkage from participants. This contract will include agreed wording of the consent form which is approved by HIPE to ensure it addresses any data protection issues and sufficiently covers any confidentially assurances made by HIPE to patients at the time of collection.</td><td><b>Outcome:</b> The suggested solution will ensure that any concerns in relation to the consent process are addressed in advance of contact with the participants and therefore eliminate the privacy risk.</td></tr>
</table>

<table>
<tr><td colspan="2"><b>Risk 2:</b> Both parties will have personal data disclosed to them, of which they are not the data controllers. There is a risk that information will be used inappropriately or disclosed further to additional parties.</td></tr>
<tr><td><b>Proposed solution:</b> A legally binding contract between SLS and HIPE will be developed and agreed in advance of seeking consent for linkage from participants. This contract will include clauses on the obligations of both organisations in relation to confidentiality of the shared data, protecting the privacy of the participants, approved uses and disclosure of the shared data and retention and destruction standards.</td><td><b>Outcome:</b> While the proposed contract will not fully eliminate the risk of inappropriate use or disclosure, it will significantly reduce both the likelihood and consequences of it occurring and enables the risk to be managed using legal means. There will be financial penalties for breach of contract by either organisations.</td></tr>
<tr><td colspan="2"><b>Risk 3:</b> Not providing sufficient data from SLS to HIPE may result in inefficient matching and requesting insufficient data from HIPE to SLS may limit the research that can be conducted on the linked data. Both of these scenarios may require the linkage process be repeated which increases the opportunity for error or privacy risk.</td></tr>
<tr><td><b>Proposed solution:</b> In advance of beginning the project a consultation process will take place between SLS and HIPE. Discussions from this forum will ensure HIPE receive all relevant variables required for effectively identifying SLS participants in the HIPE database. Similarly, a review of available HIPE data in conjunction with SLS's research aims will be conducted to ensure all required data is requested without being excessive or beyond the remit of SLS research.
Once variables have been agreed on through this consultation process, they will be documented in a Data Transfer Agreement which will be included in the SLS-HIPE contract.</td><td><b>Outcome:</b> All variables necessary for HIPE to identify SLS participants as well as the HIPE data which SLS will receive about the matched participants will be agreed in writing, eliminating the risk of error or oversight in this process.</td></tr>
</table>

<table>
<tr><td colspan="2"><b>Risk 4:</b> If an inadequate linkage process is conducted there is a risk of incorrect matching which will result in incorrect information being assigned to participants. There is also a risk that details of participants who did not consent to inclusion in the project will be incorrectly sent to HIPE.</td></tr>
<tr><td><b>Proposed solution:</b> While incorrect linkage may represent a more significant risk to research quality, there is also a risk for privacy and data protection as participants have a legal right for adequate and complete data. An agreed data linkage protocol will be developed in advance. This will establish what constitutes a successful match and set standards for degrees of disagreement between the SLS and HIPE data at which a match will not be accepted. This is protocol will also include safety checks which will ensure only participants who consent to linkage are included in the dataset sent from SLS to HIPE.</td><td><b>Outcome:</b> Complete eradication of data linkage error is not possible, particularly in the absence of an established unique identifier across the SLS and HIPE datasets. However, a defined linkage protocol will minimise the level of error and therefore reduce the risk to an acceptable level. This minimised risk of data linkage error is not so high as to prevent the project from proceeding as the risk is inherent in all data linkage and the potential benefits outweigh the outstanding risk.</td></tr>
<tr><td colspan="2"><b>Risk 5:</b> Risks associated with responsibility in relation to data access requests. SLS will need to disclose the HIPE data to the participant which it relates to if a subject access request is received.</td></tr>
<tr><td><b>Proposed solution:</b> If a subject access request is received from a participant in relation to the data held by SLS, they will be legally required to disclose all data to the participant including information obtained though the HIPE data linkage. Disclosure under these circumstances will be included as an agreed use in the SLS-HIPE contract. The source of the data will be identified as HIPE. The SLS-HIPE contract will also set processes for data correction requests. Any data correction requested by the participant in relation to HIPE data held by SLS will be updated in the SLS datasets and the participant will be informed to contact HIPE in relation to correction of the original data.</td><td><b>Outcome:</b> An agreed procedure for processing data subject access requests will be included in the SLS-HIPE contract eliminating any ambiguity in either SLS or HIPE responsibility.</td></tr>
</table>

Recommendations

This PIA was conducted to identify and address the potential privacy risks of a linkage project between SLS and the HIPE system. Of the five risks identified during Stage 2, sufficient solutions have been proposed for each of them through the review conducted in Stage 3. These solutions have been incorporated into the recommendations below which will serve to mitigate the risks associated with the project and maintain the privacy rights of included participants:

- A legally binding contract between SLS and HIPE will be developed and agreed in advance of seeking consent for linkage from participants. This contract will detail in full the obligations of both organisations in relation to confidentiality of the shared data, protecting the privacy of the participants, approved uses of the shared data, approved disclosure including processing of subject access requests and retention and destruction standards.

- Information booklets and consent forms will be developed specifically for the project and will be approved by an academic ethics committee, a data protection lawyer, SLS and HIPE in advance of seeking consent from participants.

- Based on consultation between SLS and HIPE, a Data Transfer Agreement (DTA) will be developed which includes details on all data to be transfers between SLS and HIPE. Only data detailed in this DTA will be shared between the two organisations.

- A set data linkage procedure will be agreed in advance of any matching attempts. This will set out matching standards, minimising linkage errors and therefore maximise the likelihood that a participant identified in the HIPE dataset truly refers to the same person in the SLS dataset.

Appendix C: Survey of Longitudinal Researchers – Questionnaire

Online version available at:

https://scsstcd.qualtrics.com/jfe3/preview/SV_5BYos8Z6i5LcwLP?Q_CHL=preview

Question 1: **What is your career stage?**

1. Undergraduate
2. Postgraduate
3. Postdoctoral
4. Mid-career
5. Senior

Question 2: **What are the primary focus areas of your research? (Select all that apply)**

1. Pregnancy and maternal health
2. Childhood health and experiences
3. Ageing
4. Labour force dynamics
5. Health services research & health policy
6. Disability & carers
7. Disease specific research (Alzheimer's, cancer, etc.)
8. Other, please specify

Question 3: **What is your primary research area?**

1. Economic, social and behavioural science
2. Arts and humanities
3. Medical
4. Natural environment
5. Biotechnology and biological sciences
6. Engineering and physical sciences
7. Methodology/Statistics
8. Other, please specify __________________

Question 4: **In your opinion, what are the potential benefits of administrative data linkage with research data?**
(Select all that apply)

1. Reduces respondent burden
2. Data correction
3. Reduces cost of data collection
4. Data enrichment
5. Minimise the effect of attrition
6. Reduce recall bias
7. Reduce measurement error
8. Enables research that would otherwise not be possible
9. Other, please specify ___________________
10. ⊗ No benefits to administrative data linkage

IF Q4 ≠ 10 (no benefits):

Question 4a: **Please rank your selected benefits in order of important**

[Options endorsed in Q4 are displayed and can be ranked by dragging into preferred order]

Question 5: **Have you ever attempted to link administrative data to your own research data or research data you have used?**

1. Yes
2. No

IF Q5 = YES:

Question 6: **Were you able to successfully link the data?**

1. Yes
2. No

IF Q6 = YES:

Question 7: **What administrative data did you link with your research data?**

[FREE TEXT]

IF Q6 = NO:

Question 8: **What administrative data did you attempt to link to your research data?**

[FREE TEXT]

IF Q5 = YES:

Question 9: **What barriers did you encounter during this linkage process? (Select all that apply)**

1. Obtaining appropriate consent
2. Lack of unique personal identifier across administrative data
3. Ethical considerations
4. Legislative considerations
5. Technology limitations
6. Privacy concerns
7. Administrative data owners' willingness to share data
8. Format of administrative data unsuitable for sharing/reuse
9. Other, please specify_______________
10. ⊗ No barriers experienced

IF Q6 = NO & Q9 ≠ 10 (no barriers):

Question 9a: **Of the selected barriers, what do you think was the main reason you were unable to link the data?**

[Options endorsed in Q9 are displayed and single option can be selected]

Question 10: **Would you see a benefit for your research to linking any of the following administrative datasets with your existing research data?**

1. National Perinatal Reporting System
2. Vital Statistics - Death Registration
3. National Cancer Screening Service
4. National Cancer Registry Ireland
5. Primary Care Reimbursement Service
6. Hospital In-Patient Enquiry (HIPE)
7. National Hip Fracture Database
8. Irish National Orthopaedic Register
9. Irish National Pacemaker Register
10. Irish Childhood Diabetes National Register
11. National Immunisation Database
12. Primary/Post-Primary Pupils Databases
13. Child Benefit Register
14. CSO Small Area Population Statistics
15. Other, please specify

If Q10 = 15 (other)

Question 10a: **What other administrative data would you link to link with your research data?**

[Free text]

Question 11: **What do you think would be the most important addition in Ireland to facilitate future administrative linkage projects in your research area?**

[Free text]

Data Access, Storage, Sharing and Linkage Model

Question 12**: Are you aware of the DASSL model (Data Access, Storage, Sharing and Linkage) developed by the Health Research Board?**

1. Yes
2. No

If Q12 = YES:

Question 13: **Do you think the DASSL model would enable administrative data linkage in your research area if implemented?**

1. Yes
2. No

Health Information and Quality Authority

Question 14: **Are you aware of the Information Management Standards for National Health and Social Care Data Collectors developed by the Health Information and Quality Authority (HIQA)?**

1. Yes
2. No

Privacy impact Assessment

Question 15: **Have you ever conducted a privacy impact assessment (PIA) for your research?**

1. Yes
2. No

IF Q15 = YES:

Question 16: **Have you ever used the HIQA privacy impact assessment tool?**

1. Yes
2. No

IF Q16 = YES:

Question 17: **Did the HIQA PIA tool help to identify all challenges of data linkage in advance?**

1. Yes
2. No

IF Q17 = NO:

Question 18: **Which barriers did the PIA not identify in advance?**

[Free text]

Data Protection

Question 19: **Have you ever completed data protection training?**

1. Yes
2. No

Question 20: **Are you currently or have you ever been an Officer of Statistics as set out in the Statistics Act, 1993?**

1. Currently
2. Previously
3. Never

Question 21: **Have you made any changes to your data collection or use policies in preparation for the incoming General Data Protection Regulation in 2018?**

1. Yes
2. No

Economic and Social Research Council's Longitudinal Studies Review 2017. More details available ***here***

Question 22: **Listed below are priority areas that longitudinal data could be used to address. Please indicate how important you think each of these longitudinal research areas will be in the future.**

	Very important	Fairly important	Important	Slightly important	Not at all important
Ageing population	○	○	○	○	○
Long-term effects of childhood and adult experience	○	○	○	○	○
Demographic shifts and mobilities	○	○	○	○	○
Health and well-being	○	○	○	○	○
Equality and inequality	○	○	○	○	○
Bio-social research and genomics	○	○	○	○	○
Diversity and identity	○	○	○	○	○

Question 23: **Please indicate how important you feel each of these methodological or technological priority issues are in relation to longitudinal studies, thinking broadly across design, implementation and analysis.**

	Very important	Fairly important	Important	Slightly important	Not at all important
Data linkage	○	○	○	○	○
Attrition, non-response and bias	○	○	○	○	○
Online and digital forms of data collection	○	○	○	○	○
Sampling and population representation	○	○	○	○	○
Design of questions, scales and measures	○	○	○	○	○
Complex analysis and modelling	○	○	○	○	○
Comparability and harmonisation	○	○	○	○	○
New forms of data collection	○	○	○	○	○
Missing data and reliability	○	○	○	○	○
National infrastructure supporting longitudinal studies	○	○	○	○	○
Biological specimens and methods of collection	○	○	○	○	○
Documentation and dissemination of data	○	○	○	○	○
Access to longitudinal data	○	○	○	○	○
Mixed mode data collection	○	○	○	○	○

Question 24: **Do you have any further comments on any of the issues covered (or not covered) in this survey?**

[Free text]

Thank you for taking the time to complete this survey.

By selecting the 'Next' button your answers will be submitted

If you would like to exit without submitting, please close the window

Appendix D: Survey of Longitudinal Researchers – Information Sheets

TRINITY COLLEGE DUBLIN

INFORMATION SHEET FOR PROSPECTIVE STUDIES

Project: The opportunities and challenges of linking health and administrative data with research data: A case-study review of using data linkage with longitudinal surveys

I would like to invite you to participate in a research study entitled "The opportunities and challenges of linking health and administrative data with research data: a case-study review of using data linkage with longitudinal surveys" which is being undertaken as part fulfilment for a MSc in Health Informatics in Trinity College Dublin, the University of Dublin, Ireland

Participation in this study is entirely voluntary and participants may withdraw consent to participate at any time, without need for explanation or penalty.

Name of the researcher: Margaret Foley

Purpose of the research:
The purpose of this research is to explore how data linkage can be facilitated in Ireland while continuing to protect citizens' right to privacy. The information provided by your research team will contribute to this by identifying existing examples of data linkage being undertaken and assess the demand for further potential linkage projects.

Why has this longitudinal study been asked to take part?
As an established longitudinal study in Ireland, I would greatly appreciate the opportunity to include the research team in this survey so their experiences can help inform the objectives of this research.

What is involved in in including our research team in this survey?

If you would like your longitudinal research team to be included in this research, I would require access to the team's email list. The provided emails would only be used for initial contact and one reminder in relation to taking part in the survey.

Participation of individual researchers in the survey is entirely voluntary. Once they have consented to taking part, they will be asked a series of questions in relation to data linkage between administrative and longitudinal survey data for their research projects.

Who will have access to the information our research team provides?

All the information provided will be kept strictly confidential. No contact information will be collected during the questionnaire meaning it will not be possible to link individual answers back to participants. Only aggregated result will be used for publications and no individual participant will be named. Disaggregated data will only be available to the lead researcher and the research supervisors.

The provided research team email list will be only be available to the lead researcher and will be deleted on completion of the survey fieldwork period.

Who is coordinating this research?

The research is being coordinated by Margaret Foley as lead researcher. Additionally, supervisors from Trinity College Dublin are overseeing and guiding the research. No funding has been received and there are no external parties involved in coordinating the study.

How will the information be protected?

All data collected will be stored and processed in accordance with the Data Protection Acts 1988 and 2003. All data will be stored securely in password protected files. No contact information will be collected during the questionnaire and data will be anonymised. All published results will be based on aggregated data. Please note, in the extremely unlikely event that illicit activity is reported during the questionnaire, it will be reported to the appropriate authorities. One category of illicit activity that responses to this questionnaire may reveal is Data Protection violation. If such is identified, the Data Protection Commissioner is the appropriate authority. Details of offences and penalties are listed by the office of the Data Protection Commissioner:

https://www.dataprotection.ie/docs/Offences-and-Penalties-under-the-Data-Protection-Act/r/97.htm

How will the data provided be used?

The research team email list will only be used to distribute survey information and a link for potential participants to provide consent and take part in the survey. The list email list will also be used to send a reminder to participants two weeks after initial contact. The results of the questionnaire will be presented as part of my dissertation and submitted to Trinity College Dublin. The results may also be used for presentations at conferences, seminars or workshops or submitted to peer-reviewed journals. In all cases only aggregated results will be published.

Additionally, once the dissertation has been completed, I will make aggregated results available to any research participant upon request by email (mfoley1@tcd.ie).

Does this study have ethical approval?

Ethical approval has been received from the Research Ethics Committee of the School of Computer Science and Statistics, Trinity College Dublin.

What if I have further questions?

If you have any questions before, during or after completion of this questionnaire please do not hesitate to contact me on email (mfoley1@tcd.ie) or phone (085-7515773). Participants can also contact me on these details if they require a debrief on completion of the questionnaire.

Conflict of interest:

The researcher works in the area of longitudinal surveys, however, as all answers will be anonymised it will not be possible for the researcher to identify participants from the information they provide.

TRINITY COLLEGE DUBLIN

INFORMATION SHEET FOR PROSPECTIVE PARTICIPANTS

Project: The opportunities and challenges of linking health and administrative data with research data: A case-study review of using data linkage with longitudinal surveys

I would like to invite you to participate in a research study entitled "The opportunities and challenges of linking health and administrative data with research data: a case-study review of using data linkage with longitudinal surveys" which is being undertaken as part fulfilment for a MSc in Health Informatics in Trinity College Dublin, the University of Dublin, Ireland

Your participation in this study is entirely voluntary and you may withdraw consent to participate at any time, without need for explanation or penalty. Completion of each question is voluntary and you may skip any questions you wish.

Name of the researcher: Margaret Foley

Purpose of the research:
The purpose of this research is to explore how data linkage can be facilitated in Ireland while continuing to protect citizens' right to privacy. The information you provide will contribute to this by identifying existing examples of data linkage being undertaken and assess the demand for further potential linkage projects.

Why have I been asked to take part?
You have been invited to participate due to your role with a longitudinal survey in Ireland. Your experiences will help inform the objectives of this research. Your contact details were obtained either from your study website or from a senior member of your study team.

What is involved in taking part?

If you would like to take part in this research, you will first need to consent to participation on the next page (select Next button). Once consent has been received you will be asked a series of questions in relation to data linkage between administrative and longitudinal survey data. This questionnaire will take approximately 10-15 minutes to complete.

Who will have access to the information I provide?

All the information you provide will be kept strictly confidential. No contact information will be collected during the questionnaire meaning it will not be possible to link individual answers back to participants. Email addresses used for initial contact will be deleted on completion of the questionnaire fieldwork period. Only aggregated results will be used for publications and no individual participant will be named. Disaggregated data will only be available to the lead researcher and the research supervisors.

Who is coordinating this research?

The research is being coordinated by Margaret Foley as lead researcher. Additionally, supervisors from Trinity College Dublin are overseeing and guiding the research. No funding has been received and there are no external parties involved in coordinating the study.

How will my information be protected?

All data collected will be stored and processed in accordance with the Data Protection Acts 1988 and 2003. All data will be stored securely in password protected files. No contact information will be collected during the questionnaire and data will be anonymised. All published results will be based on aggregated data. Please note, in the extremely unlikely event that illicit activity is reported during the questionnaire, it will be reported to the appropriate authorities. One category of illicit activity that responses to this questionnaire may reveal is Data Protection violation. If such is identified, the Data Protection Commissioner is the appropriate authority. Details of offences and penalties are listed by the office of the Data Protection Commissioner:

https://www.dataprotection.ie/docs/Offences-and-Penalties-under-the-Data-Protection-Act/r/97.htm

How will the data I provide be used?

The results of the questionnaire will be presented as part of my dissertation and submitted to Trinity College Dublin. The results may also be used for presentations at conferences, seminars or workshops or submitted to peer-reviewed journals. In all cases only aggregated results will be published.

Additionally, once the dissertation has been completed, the lead researcher will make aggregated results available to any research participant upon request by email (mfoley1@tcd.ie).

Does this study have ethical approval?

Ethical approval has been received from the Research Ethics Committee of the School of Computer Science and Statistics, Trinity College Dublin.

What if I have further questions?

If you have any questions before, during or after completion of this questionnaire please do not hesitate to contact me on email (mfoley1@tcd.ie) or phone (085-7515773). Please contact me on these details if you require a debrief on completion of the questionnaire.

Conflict of interest:

The researcher works in the area of longitudinal surveys, however, as all answers will be anonymised it will not be possible for the researcher to identify participants from the information they provide.

Appendix E: Survey of Longitudinal Researchers – Consent Form

TRINITY COLLEGE DUBLIN

INFORMED CONSENT FORM

Project: The opportunities and challenges of linking health and administrative data with research data:

A case-study review of using data linkage with longitudinal surveys

I would like to invite you to participate in a research study entitled "The opportunities and challenges of linking health and administrative data with research data: a case-study review of using data linkage with longitudinal surveys" which is being undertaken as part fulfilment for a MSc in Health Informatics in Trinity College Dublin, the University of Dublin, Ireland

Lead Researcher: Margaret Foley

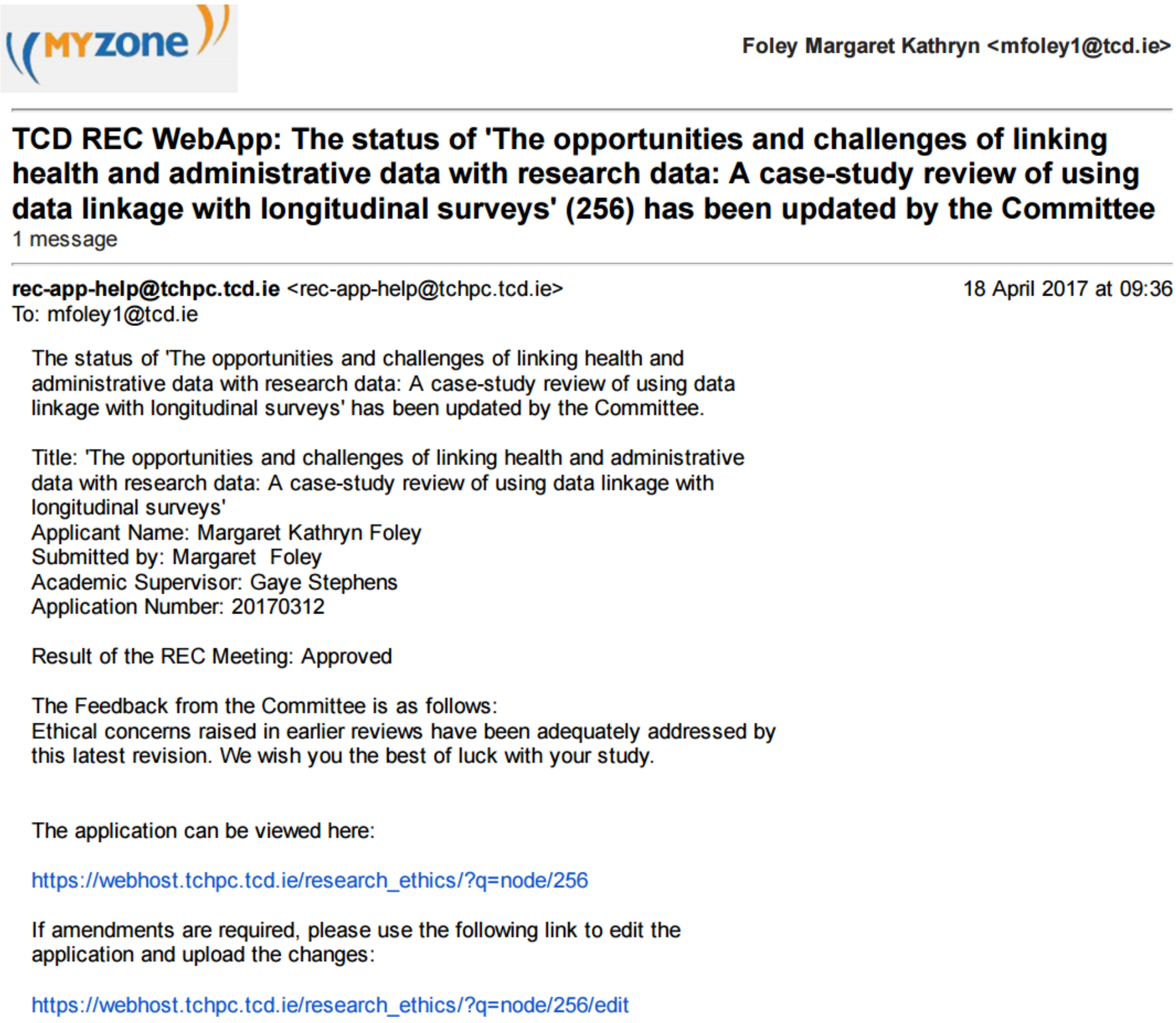
DECLARATION:

- I am 18 years or older and am competent to provide consent.
- I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to appropriate authorities.
- I understand that Data Protection violations will be reported to the Data Protection Commissioner.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- I understand that I should not name any individuals in any open fields of the questionnaire. Any such replies will be anonymised.
- As participation involves the use of a computer monitor, I understand that if I or anyone in my family has a history of epilepsy then I am proceeding at my own risk.

Statement of investigator's responsibility: I have explained the nature and purpose of this research study, the procedures to be undertaken and any risks that may be involved. I have offered to answer any questions and fully answered such questions. I believe that the participant understands my explanation and has freely given informed consent.

No, I do not consent	Yes, I consent
○	○

Appendix F: Ethical Approval Confirmation Email



Foley Margaret Kathryn <mfoley1@tcd.ie>

TCD REC WebApp: The status of 'The opportunities and challenges of linking health and administrative data with research data: A case-study review of using data linkage with longitudinal surveys' (256) has been updated by the Committee

1 message

rec-app-help@tchpc.tcd.ie <rec-app-help@tchpc.tcd.ie>
To: mfoley1@tcd.ie

18 April 2017 at 09:36

The status of 'The opportunities and challenges of linking health and administrative data with research data: A case-study review of using data linkage with longitudinal surveys' has been updated by the Committee.

Title: 'The opportunities and challenges of linking health and administrative data with research data: A case-study review of using data linkage with longitudinal surveys'
Applicant Name: Margaret Kathryn Foley
Submitted by: Margaret Foley
Academic Supervisor: Gaye Stephens
Application Number: 20170312

Result of the REC Meeting: Approved

The Feedback from the Committee is as follows:
Ethical concerns raised in earlier reviews have been adequately addressed by this latest revision. We wish you the best of luck with your study.

The application can be viewed here:

https://webhost.tchpc.tcd.ie/research_ethics/?q=node/256

If amendments are required, please use the following link to edit the application and upload the changes:

https://webhost.tchpc.tcd.ie/research_ethics/?q=node/256/edit

Appendix G: Details of Included Papers

Authors	Year of publication	Country	Title
Al Baghal et al.	2016	UK	Obtaining data linkage consent for children: factors influencing outcomes and potential biases
Audrey et al.	2016	UK	Young people's views about consenting to data linkage: findings from the PEARL qualitative study
Audrey et al.	2016	UK	Young people's views about the purpose and composition of research ethics committees: findings from the PEARL qualitative study
Boyd et al.	2014	Australia	Technical challenges of providing record linkage services for research
Brett and Deary	2014	UK	Realising health data linkage from a researcher's perspective: Following up the 6-day sample of the Scottish mental survey 1947
Brownell et al.	2013	Canada	Administrative data linkage as a tool for child maltreatment research
Carroll et al.	2016	Australia	Agreement between self-reported healthcare service use and administrative records in a longitudinal study of adults recently released from prison
Carter et al.	2010	New Zealand	Understanding the determinants of consent for linkage of administrative health data with a longitudinal survey
Carter et al.	2012	New Zealand	Differential loss of participants does not necessarily cause selection bias.
Eapen et al.	2014	Australia	"Are you available for the next 18 months?" - methods and aims of a longitudinal birth cohort study investigating a universal developmental surveillance program: the 'Watch Me Grow' study
Fredman et al.	2001	USA	Extending gerontological research through linking investigators' studies to public-use datasets
Hagger-Johnson et al.	2016	UK	Opportunities for longitudinal data linkage in Scotland

Holman et al.	2008	Australia	A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system
Hure et al.	2015	Australia	Validity and reliability of stillbirth data using linked self-reported and administrative datasets.
Husain et al.	2012	UK	HERALD (health economics using routine anonymised linked data)
Johnson et al.	2015	USA	The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility
Knies & Burton	2014	UK	Analysis of four studies in a comparative framework reveals: health linkage consent rates on British cohort studies higher than on UK household panel surveys
Knies et al.	2012	UK	Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population
MacLeod et al.	2010	UK	The Edinburgh Addiction Cohort: recruitment and follow-up of a primary care based sample of injection drug users and non-drug-injecting controls
Mars et al.	2016	UK	Using Data Linkage to Investigate Inconsistent Reporting of Self-Harm and Questionnaire Non-Response
McGhee et al.	2015	International	Taking a Long View in Child Welfare: How Can We Evaluate Intervention and Child Wellbeing Over Time?
Mostafa et al.	2016	UK	Variation within households in consent to link survey data to administrative records: evidence from the UK Millennium Cohort Study.
Mountain et al.	2016	Australia	Data linkage in an established longitudinal cohort: the Western Australian Pregnancy Cohort (Raine) Study
Sala et al.	2014	UK	Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel
Soloff et al.	2007	Australia	Enhancing longitudinal studies by linkage to national databases: Growing up in Australia, the longitudinal study of Australian children
Al Baghal et al.	2016	UK	Obtaining data linkage consent for children: factors influencing outcomes and potential biases

Appendix H: Details of Papers Excluded on Review of Full Text

Authors	Year of publication	Country	Title	Reason for exclusion
Almeida et al.	2012	Australia	Older men who use computers have lower risk of dementia	Does not discuss challenges/benefits
Anstey et al.	2014	Australia	The influence of smoking, sedentary lifestyle and obesity on cognitive impairment-free life expectancy	Does not discuss challenges/benefits
Baba et al.	2014	Australia	A longitudinal study of foot ulceration and its risk factors in community-based patients with type 2 diabetes: the Fremantle Diabetes Study	No longitudinal survey data
Booth et al.	2014	UK	Associations between objectively measured physical activity and academic attainment in adolescents from a UK cohort	Does not discuss challenges/benefits
Bruce et al.	2010	Australia	Maternal family history of diabetes is associated with a reduced risk of cardiovascular disease in women with type 2 diabetes: the Fremantle Diabetes Study	Does not discuss challenges/benefits
Caillet et al.	2015	France	Increased Mortality for Elective Surgery during Summer Vacation: A Longitudinal Analysis of Nationwide Data	No administrative data linkage
Chawla et al.	2015	USA	Unveiling SEER-CAHPS®: a new data resource for quality of care research	No longitudinal survey data
Colvin et al.	2013	Australia	Are women with major depression in pregnancy identifiable in population health data? BMC Pregnancy Childbirth	No longitudinal survey data
Cornish et al.	2016	UK	Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R	Does not discuss challenges/benefits
Davis et al.	2014	Australia	Incidence and precipitants of hospitalization for pancreatitis in people with diabetes: the Fremantle Diabetes Study	Does not discuss challenges/benefits
Davis et al.	2007	Australia	Does self-monitoring of blood glucose improve outcome in type 2 diabetes? The Fremantle Diabetes Study	Does not discuss challenges/benefits

Egan et al.	2016	UK	Proportionate universalism in practice? A quasi-experimental study (GoWell) of a UK neighbourhood renewal programme's impact on health inequalities	No administrative data linkage
Emerson & Halpin	2013	UK	Anti-social behaviour and police contact among 13- to 15-year-old English adolescents with and without mild/moderate intellectual disability	Does not discuss challenges/benefits
Ford et al.	2016	Australia	Prospective longitudinal study of testosterone and incident depression in older men: The Health In Men Study	Does not discuss challenges/benefits
Gjersing & Bretteville-Jensen	2015	Norway	Are overdoses treated by ambulance services an opportunity for additional interventions? A prospective cohort study	Does not discuss challenges/benefits
Gopinath	2015	Australia	Age-related macular degeneration and risk of total and cause-specific mortality over 15 years	Does not discuss challenges/benefits
Guhn et al.	2016	Canada	Examining the social determinants of children's developmental health: protocol for building a pan-Canadian population-based monitoring system for early childhood development	No longitudinal survey data
Guhn et al.	2016	Canada	Associations of teacher-rated social, emotional, and cognitive development in kindergarten to self-reported wellbeing, peer relations, and academic test scores in middle childhood	No longitudinal survey data
Harris et al.	2016	Australia	End of life hospitalisations differ for older Australian women according to death trajectory: a longitudinal data linkage study	Does not discuss challenges/benefits
Hart et al.	2015	Australia	Prevalence, risk factors and sequelae of Staphylococcus aureus carriage in diabetes: the Fremantle Diabetes Study Phase II	Does not discuss challenges/benefits
Hayes et al.	2016	Australia	Early childhood obesity: Association with healthcare expenditure in Australia	Does not discuss challenges/benefits
Haynes et al.	2016	United States	Bidirectional Data Collaborations in Distributed Research	No longitudinal survey data
Jones et al.	2014	UK	The growing price gap between more and less healthy foods: analysis of a novel longitudinal UK dataset	No longitudinal survey data

Jorm et al.	2012	Australia	Assessing Preventable Hospitalisation InDicators (APHID): protocol for a data-linkage study using cohort study and administrative data	Does not discuss challenges/benefits
Kamber et al.	2008	Australia	Metformin and lactic acidosis in an Australian community setting: the Fremantle Diabetes Study	Does not discuss challenges/benefits
Kazanjian	2004	Canada	Health Care Utilization by Canadian Women	Does not discuss challenges/benefits
Mitrou et al.	2010	Australia	Antecedents of hospital admission for deliberate self-harm from a 14-year follow-up study using data-linkage	No longitudinal survey data
Oliver et al.	2016	Canada	The influence of community well-being on mortality among Registered First Nations people	No longitudinal survey data
Quach et al.	2014	Australia	Primary health-care costs associated with special health care needs up to age 7 years: Australian population-based study	Does not discuss challenges/benefits
Redded et al.	2000	USA	Applications of developmental epidemiological data linkage methodology to examine early risk for childhood disability	No longitudinal survey data
Russell	2013	Australia	Adherence to dietary guidelines and 15-year risk of all-cause mortality	Does not discuss challenges/benefits
Schildcrout & Heagerty	2011	USA	Outcome-dependent sampling from existing cohorts with longitudinal binary response data: study planning and analysis	No administrative data linkage
Smith et al.	2003	UK	The ONS longitudinal study: Quality issues from 30 years of data linkage	No longitudinal survey data
Tabuchi et al.	2016	Japan	Tobacco Price Increase and Smoking Cessation in Japan, a Developed Country With Affordable Tobacco: A National Population-Based Observational Study	No administrative data linkage
Tan et al.	2013	Australia	Characteristics and prognosis of Asian patients with type 2 diabetes from a multi-racial Australian community: the Fremantle Diabetes Study	Does not discuss challenges/benefits

Tooth et al.	2012	Australia	Does government subsidy for costs of medical and pharmaceutical services result in higher service utilization by older widowed women in Australia? BMC Health Serv Res	Does not discuss challenges/benefits
Trentham-Dietz et al.	2008	USA	Health-related quality of life before and after a breast cancer diagnosis	Does not discuss challenges/benefits
van der Ven et al.	2015	Sweden	Testing Ødegaard's selective migration hypothesis: a longitudinal cohort study of risk factors for non-affective psychotic disorders among prospective emigrants	Does not discuss challenges/benefits
Westrupp et al.	2014	Australia	Community-based healthcare costs for children born low birthweight, preterm and/or small for gestational age: data from the Longitudinal Study of Australian Children	No longitudinal survey data
Yeap et al.	2013	Australia	Higher free thyroxine levels are associated with all-cause mortality in euthyroid older men: the Health In Men Study	Does not discuss challenges/benefits
Yeap et al.	2012	Australia	Higher free thyroxine levels predict increased incidence of dementia in older men: the Health in Men Study	Does not discuss challenges/benefits
Yeap et al.	2012	Australia	Associations of total osteocalcin with all-cause and cardiovascular mortality in older men	Does not discuss challenges/benefits

Appendix I: Details of Papers Excluded on Review of Titles/Abstracts

Authors	Year of publication	Country	Title	Reason for exclusion
Ackerman et al.	2005	Australia	Integrating data to facilitate clinical research: A case study	No survey data
Asaria et al.	2016	UK	Unequal socioeconomic distribution of the primary care workforce: whole-population small area longitudinal study	No survey data
Bergen et al.	2014	UK	Alcohol-related mortality following self-harm: a multicentre cohort study	No survey data
Billie et al.	2001	Norway	Two families with phenotypically different hereditary low frequency hearing impairment: longitudinal data linkage analysis	No survey data
Bopp & Minder	2003	Switzerland	Mortality by education in German speaking Switzerland, 1990-1997: results from the Swiss National Cohort	No survey data
Bouras et al.	2015	UK	Risk of Post-Discharge Venous Thromboembolism and Associated Mortality in General Surgery: A Population-Based Cohort Study Using Linked Hospital and Primary Care Data in England	No survey data
Brameld & Holaman	2005	Australia	The use of end-quintile comparisons to identify under-servicing of the poor and over-servicing of the rich: a longitudinal study describing the effect of socioeconomic status on healthcare	No survey data
Brennan et al.	2012	USA	Linking the National Cardiovascular Data Registry CathPCI Registry with Medicare Claims Data Validation of a Longitudinal Cohort of Elderly Patients Undergoing Cardiac Catheterization	No survey data
Carter et al.	2005	Australia	Non-suicidal deaths following hospital-treated self-poisoning	No survey data
Cone et al.	2012	Australia	The methodology of the Australian Prehospital Outcomes Study of Longitudinal Epidemiology (APOStLE) Project	No survey data

Crilly et al.	2014	Australia	Expanding emergency department capacity: A multisite study	No survey data
Crooks et al.	2012	UK	Defining upper gastrointestinal bleeding from linked primary and secondary care data and the effect on occurrence and 28 day mortality	No survey data
Cutajar et al.	2010	Australia	Schizophrenia and other psychotic disorders in a cohort of sexually abused children	No survey data
Derrington	2013	USA	Development of the drug-exposed infant identification algorithm (deiia) and its application to measuring part C early intervention referral and eligibility in Massachusetts, 1998-2005	No survey data
Duke et al.	2015	Australia	Long-term mortality among older adults with burn injury: a population-based study in Australia	No survey data
Duke et al.	2016	Australia	Understanding the long-term impacts of burn on the cardiovascular system	No survey data
Duke et al.	2015	Australia	Increased admissions for musculoskeletal diseases after burns sustained during childhood and adolescence	No survey data
Duke et al.	2015	Australia	Long-term Effects of Pediatric Burns on the Circulatory System	No survey data
Duke et al.	2016	Australia	Burns and long-term infectious disease morbidity: A population-based study	No survey data
Duke et al.	2016	Australia	Increased admissions for diabetes mellitus after burn	No survey data
Duke et al.	2016	Australia	Respiratory Morbidity After Childhood Burns: A 10-Year Follow-up Study	No survey data
Eisenbach et al.	1997	Israel	The Israel Longitudinal Mortality Study--differential mortality in Israel 1983-1992: objectives, materials, methods and preliminary results	No survey data
Fatovich et al.	2010	Australia	Morbidity associated with heroin overdose presentations to an emergency department: a 10-year record linkage study	No survey data
Fear et al.	2017	Australia	Burn Injury Leads to Increased Long-Term Susceptibility to Respiratory Infection in both Mouse Models and Population Studies	No survey data

Gibson et al.	2008	Australia	Exposure to opioid maintenance treatment reduces long-term mortality	No survey data
Girgis et al.	2016	Australia	Development and Feasibility Testing of PROMPT-Care, an eHealth System for Collection and Use of Patient-Reported Outcome Measures for Personalized Treatment and Care: A Study Protocol	No survey data
Gissler	2013	Sweden	Assessment of environmental health risks is feasible by secondary use of administrative registers	No survey data
Gissler & Surcel	2012	Finland	Combining health register data and biobank data	No survey data
Goswami et al.	2013	USA	Impact of an integrated intervention program on atorvastatin adherence: a randomized controlled trial	No survey data
Haak et al.	2012	USA	Creating a data infrastructure for tracking knowledge flow	No survey data
Haber et al.	2016	USA	Constructing the cascade of HIV care: methods for measurement	No survey data
Hardelid et al.	2014	UK	Contribution of respiratory tract infections to child deaths: a data linkage study	No survey data
Huynh et al.	2016	Canada	Factors Influencing the Frequency of Emergency Department Utilization by Individuals with Substance Use Disorders	No survey data
Karmel & Rosman	2008	Australia	Linkage of health and aged care service events: comparing linkage and event selection methods	No survey data
Keating et al.	2013	Australia	Pharmaceutical utilisation and costs before and after bariatric surgery	No survey data
Kinnear et al.	2011	UK	The low uptake of breast screening in cities is a major public health issue and may be due to organisational factors: a Census-based record linkage study	No survey data
Kotelchuck et al.	2014	USA	The MOSART Database: Linking the SART CORS Clinical Database to the Population-Based Massachusetts PELL Reproductive Public Health Data System	No survey data

Lane et al.	2014	USA	New linked data on research investments: scientific workforce, productivity, and public value	No survey data
Lindgren et al.	2016	USA	Asthma exacerbations and traffic: examining relationships using link-based traffic metrics and a comprehensive patient database	No survey data
McNamara & Rosenwax	2007	Australia	Factors affecting place of death in Western Australia	No survey data
Moorin & Holman	2005	Australia	Patient-initiated switching between private and public inpatient hospitalisation in Western Australia 1980 - 2001: an analysis using linked data	No survey data
Morgan et al.	2017	Australia	Incidence and Risk Factors for Deliberate Self-harm, Mental Illness, and Suicide Following Bariatric Surgery: A State-wide Population-based Linked-data Cohort Study	No survey data
Naess et al.	2013	Norway	The Norwegian Family Based Life Course (NFLC) study: data structure and potential for public health research	No survey data
Nderitu et al.	2014	UK	Analgesia dose prescribing and estimated glomerular filtration rate decline: a general practice database linkage cohort study	No survey data
Nedkoff et al.	2012	Australia	Temporal trends in the incidence and recurrence of hospitalised atherothrombotic disease in an Australian population, 2000-07: data linkage study	No survey data
O'Reilly et al.	2012	UK	Using record linkage to monitor equity and variation in screening programmes	No survey data
Reith et al.	2003	Australia	Adolescent self-poisoning: a cohort study of subsequent suicide and premature deaths	No survey data
Remy et al.	2014	USA	Longitudinal analysis of health outcomes after exposure to toxics, Willits California, 1991-2012: application of the cohort-period (cross-sequential) design	No survey data

Renzi et al.	2016	UK	Do colorectal cancer patients diagnosed as an emergency differ from non-emergency patients in their consultation patterns and symptoms? A longitudinal data-linkage study in England	No survey data
Rørth et al.	2016	Denmark	The importance of β2-agonists in myocardial infarction: Findings from the Eastern Danish Heart Registry	No survey data
Rushmer et al.	2011	UK	Is the routine recording of primary care consultations possible ... and desirable? Lessons for researchers from a consultation with multiple stakeholders	No survey data
Shepard et al.	2003	UK	Linkage analysis of cross-sectional and longitudinally derived phenotypic measures to identify loci influencing blood pressure	No survey data
Spilsbury et al.	2015	Australia	Cross border hospital use: analysis using data linkage across four Australian states	No survey data
Stender et al.	2015	Denmark	Micro data integration for Labour Market Account	No survey data
Stevenson et al.	2016	Australia	Burn leads to long-term elevated admissions to hospital for gastrointestinal disease in a West Australian population based study	No survey data
Strazdins et al.	2016	Australia	Intergenerational policy and workforce participation in Australia: using health as a metric	No administrative data linkage
Streart et al.	2015	Australia	Administrative data linkage as a tool for developmental and life-course criminology: The Queensland Linkage Project	No survey data
Tajima et al.	1998	Japan	Risk factors for liver dysfunction in middle aged men based on four year health examination data	No administrative data linkage
Thomson et al.	2006	Australia	A long-term population-based clinical and morbidity profile of Angelman syndrome in Western Australia: 1953-2003	No survey data
Tu et al.	2006	USA	Second-Order Linkage and Family Datasets	No survey data
Walsh et al.	2015	International	School-based education programmes for the prevention of child sexual abuse	No survey data

Wilson et al.	2010	USA	Application of a New Method for Linking Anonymous Survey Data in a Population of Soldiers Returning from Iraq	No administrative data linkage
Zhao et al.	2015	Australia	Assessing improvements in survival for stroke patients in the Northern Territory 1992-2013: a marginal structural analysis	No survey data