Chapter 1

Methods

This chapter outlines the process of obtaining administrative data suitable to answer the thesis research questions and has seven sections. Firstly, a brief description of administrative data linkage research and its associated advantages and disadvantages are discussed. The following section provides a detailed description of the strict information governance protocols required including: the infrastructure used, the approvals process, and the legal framework enabling the research to take place. Thirdly, a thorough description of the data sources used in analysis is provided outlining their source, reasons for collection, and any known quality issues. In the fourth section, the process of probabilistically matching personal identifiers from the Social Care Survey to a research population spine is described. This process enabled the linkage of social and health care data sources. The fifth section describes how the study cohort was created including the data wrangling procedures necessary to join together data from disparate sources. The penultimate section outlines the statistical methods applied to the data in order to answer the thesis research questions. Finally, An illustrated timeline depicting the major stages of the overall project is displayed with a brief description.

1.1 Data Linkage

> Record linkage refers to a merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record.
>
> *[OECD, 2006]*

Administrative data is data that is generated when individuals use a service of some description. Often in research terms, and exclusively in this thesis, administrative data refers to data generated by the use of *public* services (Pavis and Morris, 2015; Mazzali and Duca, 2015). This data can describe the provision of a specific service or how it was administered by the provider (Pavis and Morris, 2015; Mazzali and Duca, 2015). As the above definition outlines, record linkage involves joining data about individuals from two or more administrative databases together (Fleming, Kirby and Penny, 2012; Harron, 2016) and is being increasingly used in social science research (Atherton *et al.*, 2015; Bell *et al.*, 2016).

Using administrative data for research purposes has a number of advantages and disadvantages. The data is not collected for research purposes and as such may lack specific information relevant to a researcher's line of inquiry (Mazzali and Duca, 2015). This also reduces the ability of a study to adjust for all potential confounding variables, decreasing the ability to make causal inferences from analyses (Mazzali and Duca, 2015). There is the potential for ambiguity about the coding of variables in a database and what each code represents (Mazzali and Duca, 2015; Atherton *et al.*, 2015; Walesby, Harrison and Russ, 2017) which means specialist knowledge of the database and collection methods are required (Mazzali and Duca, 2015). Administrative databases also have the potential to contain data of questionable quality and high levels of missing data (Walraven and Austin, 2012; Hashimoto *et al.*, 2014; Harron *et al.*, 2017). Data can be missing for the same reasons as seen in other forms of research but, in addition, individuals may also be missing due to failure to interact with a service or because insufficient information was available to accurately match records during the data-linkage process (Harron *et al.*, 2017).

Advantages of administrative databases are that they enable large, often population sized, sample sizes because they are generated from service use (Mazzali and Duca, 2015; Pavis and Morris, 2015; Walesby, Harrison and Russ, 2017). This characteristic also reduces the potential for sampling bias (Mazzali and Duca, 2015). Well maintained administrative data can offer information over long periods of time including very recent data (Pavis and Morris, 2015). This can make inferences from research findings more robust with excellent levels of external validity without the extra cost traditional observational studies might incur (Mazzali and Duca, 2015; Harron *et al.*, 2017). The potential to link administrative databases from a number of sources is a significant advantage and offers insights into how services interact (Mazzali and Duca, 2015; Atherton *et al.*, 2015; Walesby, Harrison and Russ, 2017).

There are two main methods of linking data from disparate sources; deterministic matching and probabilistic matching (Fleming, Kirby and Penny, 2012; Harron, 2016; Doidge and Harron, 2018). Where differing datasets possess common unique identifiers, deterministic matching simply links data using this identifier. Probability matching methodology can be

employed in the absence of a common unique identifier (Fleming, Kirby and Penny, 2012; Harron, 2016; Doidge and Harron, 2018). Using this method, a probability of two records matching correctly is calculated based on how well the records match based on a set of common partial identifiers such as name, date-of-birth, and postcode (Fleming, Kirby and Penny, 2012; Harron, 2016; Doidge and Harron, 2018). An important consideration when using probabilistic linkage is making an assessment of false-positive match rates (Fleming, Kirby and Penny, 2012; Harron, 2016; Doidge and Harron, 2018). There are three main strategies to assist with this assessment; measuring error using "gold-standard" data (such as a validated external datasets), sensitivity analyses (comparing results across differing linkage parameters), and comparing linked and unlinked data according to characteristics (such as sociodemographic subgroups) (Harron, 2016).

Scotland is home to some of the best administrative databases in the world (Pavis and Morris, 2015). This is particularly due to the high-quality of health datasets that have been collected and maintained for over 40 years (Fleming, Kirby and Penny, 2012; Pavis and Morris, 2015). Whilst linkage of differing health datasets has become common over this period, new cross-sectoral linkages are beginning to emerge such as health and educational data (Wood *et al.*, 2013), and health and social care data (Witham *et al.*, 2015). These cross-sectoral linkages are providing new insights that have the potential to have lasting impact on policy and provision of services (Pavis and Morris, 2015; Atherton *et al.*, 2015).

1.2 Information Governance

Confidentiality of data subjects is an important consideration in any data linkage project. The benefits of administrative data linkage, outlined in section 1.1, are dependent on research being conducted in a legally and ethically competent fashion. Whilst full anonymisation would be an effective way to protect data subjects confidentiality, it is almost impossible to achieve this with individual-level data suitable for research purposes (Harron *et al.*, 2017).

As an alternative, a process involving robust approvals review, researcher training with associated responsibilities and sanctions, and safe haven settings are used to preserve data subject confidentiality (Harron *et al.*, 2017). This section outlines how this process was applied for the purposes of the data linkage completed in this thesis. Firstly, the various organisations that provide the infrastructure that enabled the data linkage to take place are briefly described. An overview of the various approvals and ethical panels is then provided, followed by the legal framework which enabled data processing to take place with a brief description of how confidentiality is maintained during the linkage process.

1.2.1 Infrastructure

> The key test for an acronym is to ask whether it helps or hurts communication.
>
> *Elon Musk*

1.2.1.1 Scottish Informatics and Linkage Collaboration

The Scottish Informatics and Linkage Collaboration (SILC) is an umbrella term for a number of support services that are available to individuals wishing to conduct research using linked administrative data (SILC, 2017c). Services include computing resources (provided by the University of Edinburgh), research and project coordination advice (provided by the electronic Data Research and Innovation Service (eDRIS)), and an indexing service (provided by the National Records of Scotland (NRS)) (SILC, 2017c). SILC currently has three partner institutions; the Administrative Data Research Centres (ADRC), the Farr Institute, and the Urban Big Data Centre (UBDC) (SILC, 2017c).

1.2.1.2 Urban Big Data Centre

Funding for this PhD was provided by the Scottish Government and the Economic & Social Research Council (ESRC). The bid for funding was won by UBDC which is based within the University of Glasgow. UBDC as an organisation is also funded by the ESRC and brings together data scientists and social scientists with research interests relevant to urban living such as; housing, transport, migration, and health (UBDC, 2017b). UBDC has six partner universities; Edinburgh, Bristol, Cambridge, Reading, Sheffield, and Illinois-Chicago.

The linkage project described in this thesis was completed with the assistance of UBDC's controlled data service. This service helps researchers to access personal data that exists in administrative databases (UBDC, 2017a). In addition to a vigorous approval process, access to data is tightly controlled via safe haven IT architecture which monitors use of data and output of analyses to ensure individual anonymity is maintained (UBDC, 2017a). UBDC arranges access to the safe haven environment though liaison with eDRIS, provided by the Information Services Division (ISD) of NHS National Services Scotland (NSS) under the auspices of SILC. A more detailed description of the safe haven is given in section 1.2.3.2.

1.2.1.3 electronic Data and Research Innovation Service

ISD is a subdivision of NHS NSS (ISD, 2010b). NSS is a national NHS board in its own right and works with the other NHS boards, particularly the 14 geographic health boards, to provide centralised services such as; procurement, legal support, IT, and public health intelligence (NSS, n.d.). As a division of NSS, ISD provides, among other things, support for the latter two of these services (ISD, 2010b). This includes administering the large number of databases containing information on health service use in Scotland varying from maternity & births, to cancer services (ISD, 2010b). ISD held databases used in this thesis, the Prescribing Information System and Unscheduled Care Data Mart, are described more fully in sections 1.3.3 and 1.3.4.

eDRIS is part of ISD and provides services under SILC (SILC, 2017b). It is detailed specifically with assisting research using health administrative datasets. Researchers using the eDRIS service have a named research assistant who provides advice on; data sources, study design, the information governance approvals system, access to the safe haven environment, and review of analysis outputs to ensure disclosive information cannot be inferred (ISD, 2010b).

1.2.1.4 National Records of Scotland

NRS collects and maintains information about the people of Scotland including births, deaths, and marriages (NRS, 2017). In addition to producing annual reports and population estimates, NRS provides the indexing service under SILC which enables anonymous linking of administrative databases as a Trusted Third Party. This is made possible using an indexing spine which is based on the NHS Central Register (NHSCR) and held by NRS (NRS, 2018a). This is described more fully in section 1.4 .

1.2.1.5 Health and Social Care Analysis Division

The Health and Social Care Analysis Division (HSCAD) is a division within the Scottish Government that provides statistic, economic, and research evidence to inform policy making in this area (Scottish-Government, 2017a). It is one of many Analytical Service Divisions (ASD) that provide analytical support, advice, and briefing to the Government. HSCAD creates reports and publications in a number of key areas including; social care, care homes, and mental health and is responsible for collecting and publishing the Social Care Survey (SCS) described in section 1.3.1.

1.2.2 Approvals

As described above, one of the ways in which data subject confidentiality is maintained in data linkage projects is through a rigorous and robust approvals process. Three separate approvals were required for the purposes of data linkage in this thesis.

1.2.2.1 Research Approvals Committee

Data linkage for the project was facilitated by UBDC's controlled data service. In order to utilise this service, the research proposal required approval from the UBDC Research Approvals Committee (RAC). A full list of RAC members is available on-line (UBDC, 2017a). This committee is independent of UBDC and approves use of funds and infrastructure in UBDC and includes a lay member of the public (UBDC, 2017a). An application to use the controlled data service is judged on its academic merit, public benefit, skill of research team, and alignment with UBDC aims before being approved (UBDC, 2017c).

The approval for the main linkage project is shown in Appendix A.

1.2.2.2 Ethics

Ethical approval for data analysis was sought and gained from the University of Glasgow College of Social Sciences Research Ethics Committee (CoSS REC). A blanket ethical approval, obtained by eDRIS from the NHS East of Scotland REC, covers research that uses NHS Health data, does not involve direct contact with data subjects, has peer-review approval, stores data in the national safe haven, and is conducted by research teams based in the UK (ISD, 2010a). The main linkage project therefore only required further approval from CoSS REC to cover the non-health related data (i.e. the Social Care Survey).

The CoSS REC approval letter for the main linkage project is shown in appendix B.

1.2.2.3 Public Benefit and Privacy Panel for Health & Social Care

In addition to RAC and College ethical approval, the main thesis project also required clearance from the Public Benefit and Privacy Panel for Health & Social Care (PBPP). This was because data from NHS sources were being used. The PBPP acts as a decision making body with delegated responsibility from NHS Scotland Chief Executive Officers and the Registrar General (Scottish-Government, n.d.). Using terms of reference and guiding principles, the panel adjudicates whether research projects using administrative data generated by the NHS in Scotland can be used for research purposes. The panel

ensures that the basis for disclosing data has a clear public benefit and ensures the legal framework for accessing and processing data is sound.

The approval letter for the main thesis project is shown in Appendix C. A full description of data processing including its legal basis is presented in section 1.2.3.

1.2.3 Data processing

1.2.3.1 Legal framework

The permissions and linkage of data for this project were completed in advance of the European Union (EU) General Data Protection Regulation (GDPR) coming into effect in May 2018. The information governance was informed by antecedent laws including the Data Protection Act (DPA) (1998). However, as the study period was known to overlap with the implementation of GDPR, all legal documentation was completed to ensure compliance with the incoming regulation.

Data sharing and processing can be completed without consent of data subjects as long as certain criteria, explicitly named in legislation, are met (Bell *et al.*, 2016). For the purposes of this thesis fair processing of data was completed, without consent, in accordance with three legislative paragraphs:

- Schedule 2:(6) of the DPA.

  1. The processing is necessary for the purposes of legitimate interests pursued by the data controller or by the third party or parties to whom the data are disclosed, except where the processing is unwarranted in any particular case by reason of prejudice to the rights and freedoms or legitimate interests of the data subject.

- Schedule 3:(8) of the DPA (emphasis added)

  1. The processing is necessary for medical purposes and is undertaken by

     (a) a health professional, or

     (b) a person who in the circumstances owes a duty of confidentiality which is equivalent to that which would arise if that person were a health professional.

  2. In this paragraph "medical purposes" includes the purposes of preventative medicine, medical diagnosis, *medical research*, the provision of care and treatment *and the management of healthcare services*

- Paragraph 9 of the Data Protection (Processing of Personal Data Order 2000(SI 2000 No.417)).

  The project has clear and substantial public interest in the information it will provide to inform the delivery of public services. The data processing is necessary to enable this research to take place. The project does not support measures of analysis with respect to any individual. Finally, the project will not cause any substantial damage or distress to any individual.

Lawful processing of data for the purposes of the project is in accordance with a further two legal acts:-

- Social Work (Scotland) Act 1968.

  8 Research

  1. The Secretary of State may conduct or assist other persons in conducting research into any matter connected with his functions or the functions of local authorities in relation to social welfare, and with the activities of voluntary organisations connected with those functions.
  2. Any local authority may conduct or assist other persons in conducting research into any matter connected with their functions in relation to social welfare.
  3. The Secretary of State and any local authority may make financial assistance available in connection with any research which they may conduct or which they may assist other persons in conducting under the provisions of this section.

- National Health Service (Scotland) Act 1978

  47 Education and research facilities. (2)Without prejudice to the general powers and duties conferred or imposed on the Secretary of State under the Scottish Board of Health Act 1919, the Secretary of State may conduct, or assist by grants or otherwise any person to conduct, research into any matters relating to the causation, prevention, diagnosis or treatment of illness, or into such other matters relating to the health service as he thinks fit."

Information governance for the project was also informed by the "Guiding Principles for Data Linkage" report produced by the Scottish Government (2012). These guidelines, themselves informed by legislation such as the DPA, highlight the importance of public interest, transparency, and privacy when conducting data linkage projects with publicly held data sets.

In order to preserve anonymity whilst linking administrative data from different agencies, a method known as "linkage using a separation of functions" is employed (Pavis and Morris, 2015; Harron, 2016). This process involves the use of a Trusted Third Party (TTP) to process non-anonymised information in order to link more than one dataset together. The TTP receives personal information (e.g. names, addresses and dates-of-birth) from the data controllers of the administrative datasets to be used and creates index "keys" to send back to the data controllers to attach to their data (Pavis and Morris, 2015; Harron, 2016). The TTP creates a lookup table of index "keys" relevant to each dataset and sends these to a linkage agent. The linkage agent receives data from the data controllers *without* personal information and links them together using the "keys" created by the TTP and makes this available to a researcher in a secure environment (Pavis and Morris, 2015; Harron, 2016). This process means the TTP receives lots of personal information but no other information, the researcher has access to information relevant to their study but no personal information, and the data controllers share information about individuals in their datasets without compromising anonymity and without seeing data from each others databases (SILC, 2017a).

1.2.3.2 Safe Haven environment

Another integral part of ensuring the confidentiality of data subjects within large, linked administrative data is by holding such data in a safe haven environment (Harron *et al.*, 2017). As described in section 1.2.1.2, access to data for this thesis was administered via UBDC's controlled data service and further liaison with eDRIS to enable use of the NSS National safe haven. All data shared for the purposes of the thesis was transferred by data controllers to the safe haven by secure file transfer protocol.

The safe haven enables secure data storage and access via a Virtual Private Network (VPN) connection with strict access control. This environment does not enable external access of any kind i.e internet or saving & printing facilities (ISD, 2010c). In order to retrieve output of analyses, work was submitted for statistical disclosure control which was conducted by eDRIS employees. This process ensures that data taken out of the safe haven cannot be used, either on its own or by being combined with other data, to breach the privacy of any individual (ISD, 2010c; Harron *et al.*, 2017). A full guide to statistical disclosure control is provided by Lothian & Ritchie (2017).

1.2.3.3 Data sharing agreement

For the purposes of the main linkage project, a three-way data sharing agreement (DSA) between the University of Glasgow, NHS National Services Scotland, and Scottish Ministers (Scottish Government) was signed. This detailed the purpose of data sharing, as well

as the transfer, protection, and security of data. The roles and responsibilities of each organisation in relation to relevant data protection legislation are clearly detailed in the DSA which is shown in Appendix D.

1.3 Data Sources for Linkage

Research conducted with administrative data requires a thorough description of databases used (Walraven and Austin, 2012). This should include a description of the purpose of the data collection and the methods employed to collect data. This enables appraisal of any potential biases that may exist within the databases (Walraven and Austin, 2012; Mazzali and Duca, 2015). There are 4 main sources of data used in the main analyses of this thesis: the Social Care Survey, the National Records of Scotland population spine and death records, the Prescribing Information System, and the Unscheduled Care Data Mart. Each of these are described in more detail below.

1.3.1 Social Care Survey

The Social Care Survey (SCS) is collected annually by HSCAD for the Scottish Government to provide descriptive statistics of the amounts of social care delivered by each of Scotland's 32 local authorities (Scottish-Government, 2017b). Results are collated and published annually by HSCAD in the "Social Care Services, Scotland" report (Scottish-Government, 2017b). The SCS reports provide an overview of social care services for the public and policy makers. In addition, certain measures captured by the SCS are used in funding formulae to calculate allocation of resources to each local authority (e.g. number of people receiving home care) (Scottish-Government, 2016a).

All 32 Scottish local authorities collect information on social care as part of their management systems (Scottish-Government, 2016c). HSCAD produce a data specification document outlining the information that should be returned for the SCS and this is sent to HSCAD via a secure web-based system called ProcXed. This system supports data validation checks on transfer to improve data accuracy (Scottish-Government, 2016c). The SCS contains unique ID numbers generated by local authorities but does not routinely collect CHI numbers. In order for the SCS to be linked to health data sources, HSCAD commissioned work to link it to the NRS population spine using probabilistic linkage techniques. This process is described more fully in section 1.4.

The SCS has been collected in its present form since 2013 as a combination of two previous data collections - the Home Care Census and the Self Directed Support (Direct Payments) Survey (Scottish-Government, 2016c). Individual-level data has been collected since 2010.

Some questions have remained constant throughout this period but there have also been some changes in definitions and measures.

The most recent SCS (2017) collected information on on all individuals that received community alarm, telecare, self directed support (SDS) or social work/ support worker services during the previous financial year. In addition, individuals that received home care services, meals, housing support, shopping, or laundry services during a specified census week are included in the survey.

Before 2013, information on telecare and community alarm services was only collected for individuals receiving these services during the census week. Furthermore, the value for total weekly hours of housing support services was included in the value of home care services. The separation of these services acknowledges that housing support is often regarded as 24-hour-a-day-7-day-a-week service (Scottish-Government, 2016c).

For the value of total weekly hours of home care, HSCAD request details on scheduled and actual hours of care delivered. Some local authorities are able to return both values, others only return one value. Where both are returned, actual hours of home care are used in official reporting (Scottish-Government, 2016c). Approximately 129,000 people received community alarms and/or telecare services, approximately 60,000 received home care services, and approximately 8,000 received SDS funding in 2016/17 (Scottish-Government, 2016c). The overlap of individuals who receive more than one of these services is unknown.

The cross-sectional nature of the survey, and in particular the census week variables, mean that the SCS only collects a sample of the entire population that receive social care in Scotland in any given financial year. It is unknown how large this sample is. It is also impossible to infer whether the values of total weekly hours of home care delivered to individuals is representative of the care they receive throughout the financial year. In order to gain a better understanding of these issues, an exploratory project, using a richer data set, was conducted as part of this thesis and is reported in chapter **??**.

1.3.2 NRS population spine and Death records

The 'Research population spine' is a copy of the National Health Service Central Register (NHSCR) and is controlled by NRS (NRS, 2018a). The spine contains over 9 million records and is updated quarterly. There are very strict policies which govern access to the offline secure server where the spine is stored.

The NHSCR is used operationally for the purpose of transferring GP records (NRS, 2018a). Despite the name and the fact it is used for transferring patients records, the NHSCR does not hold medical records themselves. It contains records of anyone who was born in Scotland, who registers with a GP in Scotland, or who dies in Scotland. Variables included

are: forename, middle name, surname, date of birth, gender, postcode, and country of birth.

This resource is extremely useful for administrative data linkage projects. Data sources that don't contain a CHI number (such as the SCS) can be matched to the population spine using deterministic and probabilistic methods (see section 1.1). This then allows linkage to CHI-based data sources enabling cross-sectoral projects, such as the main analyses in this thesis, to be possible.

NRS also collates the register of deaths which includes details of every death in Scotland since 1855 (NRS, 2018b). Details on the date of death were requested for all individuals in the thesis study cohort.

1.3.3 Prescribing Information System

The Prescribing information system (PIS) contains all community prescribed medicines for every individual in Scotland from 2009 onwards. Data is collected to provide payment to community pharmacies for the medicines dispensed to the population. The data base can be linked to health sources via the Community Health Index (CHI) number. A full description of the PIS database and its applicability to research has been published by Alvarez-Madrazo et al (2016).

1.3.4 Unscheduled Care Data

The Unscheduled Care Data Mart (UCD) is a database collated by ISD for the purpose of understanding a patient journey through emergency and urgent care services (ISD, 2017). It is a linkage of routine health data from a number of sources controlled by ISD: NHS24 telephone triage service, Scottish Ambulance Service (SAS), primary care out-of-hours services (PC OOH), Accident & Emergency (A & E), acute emergency inpatient admissions (both general and mental health), and deaths. Data is available from 2011 with the exception of PC OOH data which is available from 2014. CHI numbers are available on all records.

Continuous Urgent Care Pathways (CUPs) are calculated that join together records from each of these sources that occur within 24 hrs of each other (or for services occurring within 48hrs of an acute emergency admission)(ISD, 2017). Details of all variables listed in UCD are available in a background paper published by ISD (2017). In addition to service use, UCD flags presence of any of 14 long-term health conditions in any of the above datasets and, additionally, any acute admission from 1981 onwards.

1.4 Making social care data available for linkage

The SCS does not routinely include CHI number as part of its annual data collection but does have fields for other personal identifiable information (PII) such as name, date-of-birth, gender, and postcode. In order to make the SCS available for linkage to health sources, HSCAD commissioned work to probabilistically match these identifiers to the NRS population spine (described in section 1.3.2) and create read-through indexes for linkage purposes. This work has not been published but a short report was produced.

The report described the variable quality in PII provided to the SCS by local authorities. One local authority, Clackmannanshire, returned only month/year of birth and truncated postcode data to the SCS. This meant only 1% of its records could be matched to the population spine. For this reason it was recommended that records for Clackmannanshire not be included in any analyses.

Of the remaining 31 local authorities, 17 returned date-of-birth data where a disproportionate number of records had the *day* of birth recorded as "01". Therefore these council areas had their records matched separately from the other 14 areas using a refined matching algorithm.

Using this approach an overall linkage rate of 91.2% for 31 local authorities (removing Clackmannanshire) was achieved to the population spine. Sensitivity analysis revealed fairly consistent match rates across age, sex, and SIMD deciles. However, there was much more variation in match rates at the local authority level which ranged from 76.7% to 97.9% as shown in table 1.1.

The variation in linkage rates indicates non-random missing data for SCS data derived from population spine indexes. This makes comparison of receipt of social care across local authority areas complex and a national comparison is not possible. One potential way of creating meaningful comparisons is to create sub-groups of local authorities by similar match rates and compare receipt of care within these sub-groups.

Table 1.1: Local authority linkage rates to NRS population spine

Local Authority	**Linkage rate of SCS records to NRS population spine (%)**
Angus	98.5
Dumfries & Galloway	98.5
Falkirk	97.9
Inverclyde	97.2
Argyll & Bute	96.9
South Lanarkshire	96.9
East Ayrshire	96.8
North Ayrshire	96.6
Stirling	96.5
East Renfrewshire	95.7
Glasgow City	95.7
Shetland Islands	95.5
South Ayrshire	95.4
Eilean Siar	95.2
Fife	94.7
Perth & Kinross	94
East Dunbartonshire	93.9
Edinburgh, City of	93.8
Aberdeenshire	91.5
Orkney Islands	91.4
Moray	91
Dundee City	90.6
East Lothian	86.9
West Dunbartonshire	85.1
Scottish Borders	84.2
West Lothian	83.9
Aberdeen City	82
Renfrewshire	81.1
Midlothian	80.1
Highland	79.2
North Lanarkshire	76.7

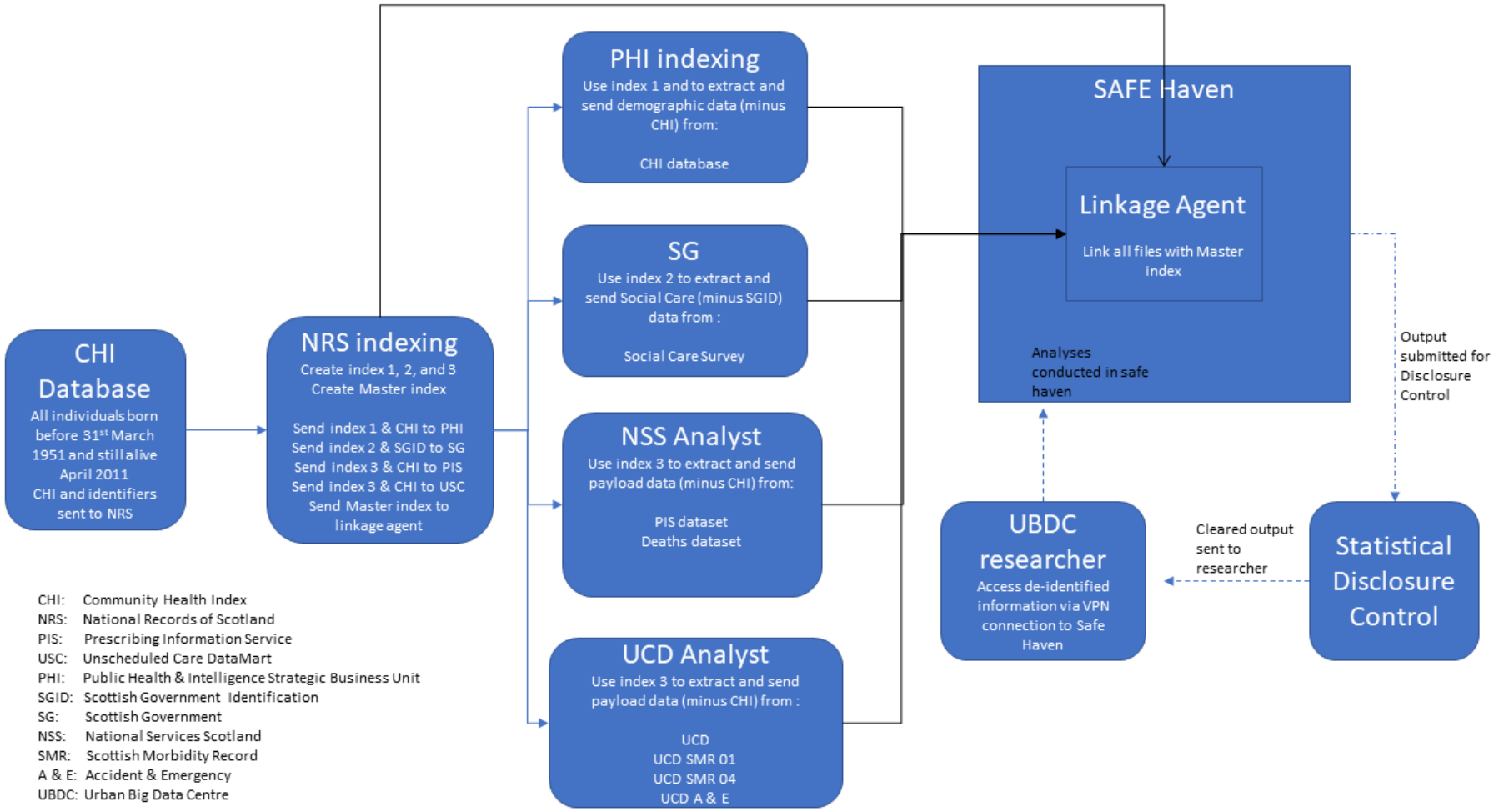
1.5 Creating a linked health and social care dataset

The study cohort included all individuals in Scotland born before 31st March 1951 and alive during the study period 1st April 2011 to 31st March 2016. This identified all those over the age of 65 (and those turning 65 during the study period). Data for the cohort was extracted from the research population spine held by NRS with CHI numbers allowing linkage to the other datasets described in section 1.3.

As figure 1.1 shows, linkage keys from the extracted cohort were sent by an eDRIS coordinator to various health and social care data sources for extraction of information relating to any of these individuals in the target data source. Specific variables requested, the time period they were requested over, and cleaning and wrangling of these data sources is described in the following section.

The aim of cleaning and wrangling was to create one row of data for each individual for each financial year (1st April - 31st March) of the study period. This format is based on the principals of tidy data (Wickham, 2014). Financial years were chosen as the time period of interest because the social care survey reports home care usage in a census week which is usually at the end of March. As each raw data file provided was in differing formats, this required differing approaches and relied heavily on data manipulation software packages `tidyr` v0.7.2 (Wickham and Henry, 2017), `dplyr` v0.7.4 (Wickham and Francois, 2017), `lubridate` v1.6.0 (Grolemund and Wickham, 2017), `stringr` v1.2.0 (Wickham, 2015), `forcats` v0.2.0 (Wickham, 2017), `data.table` v1.10.4 (Dowle *et al.*, 2018), and `zoo` v1.8-0 (Zeileis *et al.*, 2018) in the R language and environment for statistical computing version 3.4.0 (R-Core-Team, 2017) via the Integrated Development Environment RStudio v1.0.143 (RStudio-team, 2016).

Figure 1.1: Data linkage diagram



1.5.1 Demographic, geographic, and deaths information

Demographic information for all eligible individuals identified from the population spine was extracted by the Public Health and Intelligence: Strategic Business Unit at NSS. This was joined with a flag variable indicating if an individual was resident in a care home (from prescribing data) in a single file which was made available in the national safe haven. SIMD decile was assigned as per the most recent version of the area-based measure (Scottish-Government, 2016b).

Only month and year of birth were provided to avoid disclosure of identifiable information. Age was calculated by flooring each individual's **day**-of-birth to the 1st day of the **month**-of-birth provided, and then calculating the difference between this pseudo date-of-birth and the 31st of March in each financial year.

The number of observations, individuals, and the differing variables in this raw demographic file are shown in table 1.2.

Table 1.2: Demographic file data

Number of rows	Number of individuals	Variables
1,348,310	1,134,445	Index, year/month of birth, year/month of death Sex, Address start date, Address end date, Care home flag, Previous Local Authority, Current Local Authority, Current Health Board, SIMD decile

As the table indicates, some individuals had more than one row of information indicating multiple addresses during the study period (and thus potential multiple values for local, authority, health board, care home flag, and SIMD decile). To overcome this, financial year time intervals were created using the `lubridate` R package (Grolemund and Wickham, 2017). Dummy variables were then created indicating the age, local authority of residence, health board of residence, and SIMD decile during each financial year (with null values where not applicable). The variables were then gathered to long format in order to reshape the data to include one row of data per individual per financial year. Where an individual had multiple addresses during one financial year, the most recent value for local authority, health board, and SIMD decile was used. This resulted in a data frame of 7,775,410 observations pertaining to 1,134,445 individuals.

1.5.2 Social Care Survey

Data from the Scottish Care Surveys 2010 - 2016 (including separate Home Care Census and Self-Directed Support surveys for earlier years) were extracted by a Scottish Government analyst and transferred to the national safe haven in a single file. There were a number of

Table 1.3: Social Care Survey file data

Number of rows	Number of individuals	Included variables	Derived variables	Dropped variables
663,809	227,345	1. Index (ID) 2. Living alone 3. Community Alarm 4. Other telecare	1. Total weekly hours of home care 2. Home care hours group (e.g 1-5, 6-10 etc.) 3. Alarm or Telecare flag	1. Client Group 2. Eligibility Category 3. Housing Support 4. Multi Staffing 5. Scheduled Hours 6. Actual Hours

variables indicating the weekly hours of home care (if any) each individual received, whether these were provided by the local authority or an independent organisation, and whether they indicated the scheduled hours of home care or the actual number of hours delivered. There is some discrepancy between local authorities on which value (scheduled, or actual) of home care is returned to the Scottish Government for the SCS. Some authorities return scheduled only, others actual hours only, and yet others return both values. The SCS reports statistics based on the actual hours of home care delivered where available and uses the scheduled value where it is not. This convention was also used for the purposes of this thesis.

Many variables requested from the SCS had large amounts of missing data. There were also coding issues with extra values present that had no corresponding description in the provided metadata. Variables with these issues were dropped and not included in analyses. Table 1.3 lists the variables included and excluded after data cleaning.

A small fraction of observations (198 pertaining to 129 individuals) had an impossible value of weekly hours of home care greater than 168 hrs (more than 24hr-7-day-a-week care). These records were dropped from the dataset after the entire social care file had been joined to other sources of data (described in section 1.5.5).

Assessment for duplicated rows indicated 4,357 individuals had more than one row of data for some years of data. Inspection of these additional rows indicated a change in value for some variables (e.g. a flag indicating use of community alarm services positive in one row and negative in another, or different values for client group in multiple rows). These additional rows amounted to 1.1% of observations in the SCS. The exact cause of these duplications is unknown. One possible explanation is that duplication was created when records from different sources were joined together in advance of being sent for linkage. A further potential cause is the duplication of records created by the process of recycling identification numbers in some local authorities. Given the small percentage or records this affected, individuals with duplicated information were also dropped from the dataset after being joined to other files (as described in section 1.5.5).

1.5.3 Prescribing Information Service

Community prescribing information for all individuals in the cohort were extracted from the Prescribing Information System (PIS) by analysts from ISD. For each quarter of the study period (Quarter 1 2010/11 to Quarter 4 2015/16) a list of medicines prescribed to each individual was extracted and transferred in one file to the national safe haven. This file contained 134,377,877 observations of four variables: The financial year and quarter, the BNF subsection code, The approved name of the medicine, and a count of how many times the medicine was prescribed in the quarter. Coding errors were found in 138,973 observations (wrong number of digits in the BNF subsection or characters found in the count variable) and these were dropped from analysis.

The count of medicines was based on the BNF classes included in a count of polypharmacy by Guthrie et al (2015). The additional material provided on-line with this paper included a table of included drugs. This table was amended to remove BNF subsections 3.9.1, 3.9.2, and 13.9. The latter section includes different forms of shampoos whilst the former 2 sections include preparations for coughs. These were not deemed necessary to be included in overall counts. Two BNF subsections not included in the Guthrie et al table were deemed important to include as testing revealed large numbers of prescriptions included medicines from these sections would have been omitted otherwise. These sections were 2.2.4 (Potassium sparing diuretics with other diuretics) and 2.2.8. (Diuretics with potassium). In total, 198 medicines listed in the BNF were not included and rows with these medicines were removed from the PIS file. A full list of these medicines is shown in Appendix E. Table 1.4 shows the cleaning process.

Table 1.4: Data cleaning of PIS file

Reasons	Records dropped	Records remaining
Initial data file	N/A	134,377,877
Coding errors	138,973	134,238,904
Did not appear in Guthrie et al (2012) table	1,427,643	132,811,261
BNF sections 3.9.1, 3.9.2, & 13.9	645,900	132,165,361

A summary measure for each individual was created counting the total number of repeat medicines prescribed in each financial year. To be eligible in the count, a medicine had to be prescribed in at least 2 quarters of each financial year. This meant one-off prescriptions, such as antibiotics for a transient infection, were not included in the overall count. A separate count was conducted for individuals who died in the first quarter of each financial year (and thus unable to have medicines prescribed in two quarters). The total number of unique medicines prescribed in the first quarter was used for these individuals. Each participant could thus have a maximum of 6 observations, one for each financial year. A second count was created totalling the number of chapters of the BNF that each individual

had medicines prescribed from as a crude measure of body systems being treated. Table 1.5 shows the total observations and variables in the cleaned PIS file.

Table 1.5: Description of cleaned PIS file

Number or rows	Number of individuals	Variables
5,501,820	1,066,395	1. Index (id) 2. Financial Year 3. Total medicines (n) 4. Total chapters (n)

1.5.4 Unscheduled care measures

Unscheduled care information for all individuals included in the cohort was extracted from unscheduled care data mart (UCD) by an analyst from ISD. The raw file contained 3,772,402 observations from 845,893 individuals. Each observation related to a single continuous urgent care pathway (CUP) as described in section 1.3.4.

In a similar fashion to the wrangling conducted with demographics data, dummy variables were created indicating if each observation occurred within specified financial years during the study period. This enabled data to be reshaped to a long format with individuals having one or multiple rows of data for each financial year. To create one observation per individual per year, data with information on each CUP was nested within a data frame as a list column (described by Wickham & Grolemund (2017, ch.20)). With data in this format, summary measures were derived by applying functions to the list column utilising the `purrr` R package (Henry and Wickham, 2017). Derived information included counts of total USC episodes, acute admissions to hospital, A & E attendances, and total number of long-term conditions identified from admissions and A & E data. The format of the cleaned UCD data frame is described in table 1.6

Table 1.6: Description of cleaned USC file

Number of rows	Number of individuals	Variables
1,951,755	845,516	1. Index (id) 2. Year 3. USC episodes (n) 4. Acute admissions (n) 5. A & E attendances (n) 6. Long-term conditions (n)

Data were available beyond the study period ending 31st March 2016. Records outside this end date were dropped when this file was joined with the other sources of data which is described in the next section.

1.5.5 Joining sources together

Following cleaning and formatting of each individual file, further wrangling was completed which joined each file together in a parent data frame to be used for analysis. This involved loading individual files in one-by-one and joining them together using the "full join" function from the R package `dplyr` (Wickham and Francois, 2017) using the unique index number as the joining parameter. This process ensured all records were retained, even if an index number was only present in one file.

With all study data now in one data frame, further cleaning and tidying was required. This was an iterative process. As initial descriptive and statistical analysis was completed, identification of errors and data quality issues required repetition of the joining process to address these issues. This process is now described with a summary provided in table 1.7.

Table 1.7: Joining files together and cleaning process

	Number of rows	Total number of rows remaining after join/drop
Cleaned demography file	7,775,410	7,775,410
Cleaned prescribing information file	5,501,820	8,057,604
Cleaned social care file	663,809	8,072,233
Cleaned unscheduled care file	1,951,755	8,094,256
After age and death tidying process		8,115,549
Duplicates	14,809	8,100,740
Missing data for Local Authority	7,435	8,056,329
Age <65 OR Clackmannanshire OR data for 2017/18	1,832,443	6,223,886
Home care hours >168 per week	198	6,223,688
Died before 65 years of age	8808	6,214,880
Implausible SIMD value	23	6,214,857
Data from years 2010/11 OR 2016/17	1,695,602	4,519,255

Once all files had been joined together the parent data frame contained 8,094,256 observations. As there were discrepancies over time periods for which data was provided in different files, the calculation of age from the demographic file was not always present for all years of data (e.g. where demographic data was returned for the years 2015-2018 and PIS data was available from 2011-2018). To overcome this, age was recalculated from the pseudo date-of-birth (described in section 1.5.1) for all financial years. Where an individual died during a financial year, the age variable was left empty which required additional rows to be added in some cases.

As described in section 1.5.2, Approximately 4000 individuals had duplicated social care information for some years of data. These rows, and other duplicates created by the cleaning process involving age and date-of-death variables, were then dropped.

For the same reasons that age values were not shown in every year of data, values for sex, local authority of residence, health board, and SIMD decile were missing from 50,349 of observations (1.11% of the final cleaned data frame). These observations were filled by

carrying the last observation forward. Whilst this would not have affected values for sex, potential error could have be introduced to the other variables. Given the small percentage of affected observations this was deemed acceptable. Despite this, there were still 7,435 records with missing values for local authority. Cross referencing these individual rows with the raw demographics data file revealed the values for local authority in these observations were true missing data (not created by data manipulation). Given the small proportion of records these represented they were dropped from the data frame.

A further 1,832,443 observations were then removed from the data frame. These observations were for years of data where individuals were either: (a) under 65 years of age (the cohort comprised individuals over 65 or *turning* 65 during the study period), (b) resident in the Clackmannanshire local authority area (linkage rates of the social care survey to the indexing spine were too low to be reliable in this council. (See section 1.4 for details), or (c) contained unscheduled care data for financial year 2017/18 which was well beyond the study period.

Exploratory data analysis revealed three data quality issues that required further observations to be dropped from the data frame. Firstly, as described in section 1.5.2, 168 observations contained implausible values for weekly hours of home care (>168 hrs). These had not been removed as whilst cleaning the social care file so were dropped here. Secondly, calculating average age for each individual revealed 8,808 observations with a null value. Further inspection of these observations identified each individual had only one observation and had died before their 65th birthday. The inclusion criteria for the cohort stated individuals should be "born before 31st March 1951 and alive during the study period 1st April 2011 to 31st March 2016". This meant, for example, an individual born on Christmas day 1949 and dying at age 64 on Christmas day 2013 was extracted as part of the cohort data. These observations were also dropped from the parent data frame. Finally, 23 rows of data were found to have implausible values for SIMD decile. These observations were from individuals living in either the Shetland Islands or Na h-Eileanan Siar which only have data zones in 5 deciles making values outwith these deciles impossible.

Whilst the study period had been defined as 1st April 2011 to 31st March 2016 some data files contained observations outwith this period. These 1,695,602 observations were maintained for exploratory analysis but were dropped for final analysis. Thus, the final parent data frame used for all reported analyses contained 4,519,255 observations.

Derived grouping variables were created for age group (5 year bands), repeat medicines group (4 groups of similar size: 0-2, 3-5, 6-8, and over 9 repeat medicines), and linkage group (grouping councils that had linkage rates (described in section 1.4) within 4% of each other (e.g. 96-99.9%, 92-95.9% etc.)

1.6 Statistical methods

1.6.1 Research question 1

To address the question of how multimorbidity plus sociodemographic and geographic factors influence the utilisation of social care, logistic regression models were fitted separately to each financial year of data. The dependent variable in these models was receipt of any form of social care, measured by presence or not in the social care survey. Observations where an individual had died during the financial year (therefore had no chance of appearing in the SCS at the end of March) and where an individual did not receive social care but was resident in a care home (therefore not eligible for home-based social care) were excluded from the model.

Independent variables and interaction terms were added incrementally to assess impact on model fit which was measured by McFadden's pseudo R^2 (McFadden, 1974) calculated by the formula

$$R^2_{McFadden} = 1 - \frac{ln(LM_1)}{ln(LM_0)}$$

Where:

$ln(LM_1)$ = log likelihood of the fitted model and:
$ln(LM_0)$ = log likelihood of the null model (with intercept only as a predictor).

McFadden's R^2 values are not analogous with R^2 values calculated from linear models. Instead, values of 0.2 - 0.4 represent an excellent model fit (McFadden, 1977,p35; Louviere, Hensher and Swait, 2000,p55).

The final models included sex, age group, repeat medicines group, SIMD decile of residence, and local authority of residence as independent variables. Interaction terms were fitted between: sex & age group, age group & repeat medicines group, SIMD decile & repeat medicines group, and SIMD decile & local authority of residence. Exploratory models had revealed a linear effect of SIMD decile on receipt of social care thus, given the complexity of interaction terms and subsequent computational requirement, SIMD was fitted as a continuous term in the final models. As different local authorities had differing linkage rates from the social care survey to the NRS population spine (section 1.4), separate models were fitted including councils with similar linkage rates. For the purposes of this thesis, only councils with a high linkage rate of either 92-95.9% or 96-99.9% were included in models. This meant, overall, two models were fitted separately to five individual years of data resulting in 10 final models.

With the exception of SIMD decile, all independent variables were categorical in nature with a number interaction terms fitted as described above. As such, misinterpretation

of coefficients in the logistic regression model (either as odds-ratios or probabilities) was more likely (Ai and Norton, 2003; Mood, 2010; Mustillo, Lizardo and McVeigh, 2018). Therefore, estimated effects are reported as average partial effects (APEs) described by Mood (2010, p75) with the formula

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{x_1}f(\beta_{x_i})$$

Where:
β_{x_1} = the log odds-ratio for variable x_1,
β_{x_i} = the value of the logit for the i-th observation, and
$f(\beta_{x_i})$ = the probability distribution function of the logistic distribution with regard to β_{x_i}

The effect estimate describes the average marginal effect (AME) at a specific value of x_1. Williams (2012,p325) provides an intuitive example of how APEs are calculated and interpreted which has been adapted to reflect the fitted model and uses the "sex" variable as an example here

- Go to the first case. Treat that observation as if they were male regardless of actual sex. Leave other values of independent variables at their observed value. Compute the probability of receiving social care with the fitted model (including interaction terms).
- Repeat, but change the value of sex to female.
- The difference in the two probabilities is the partial (marginal) effect for that case.
- Repeat for every observation in the data.
- Compute the average of all the partial effects. This gives the APE for being female.

As Williams (2012) observes, this has the effect of comparing hypothetical populations - one female, one male - with the same observed values for other explanatory variables in the model. The only differences between these hypothetical populations is their sex with the estimate describing the differences in the probability of them receiving social care.

Where categorical variables have more than one value (e.g. age group), the APE describes the average difference in the probability between the observed value and the reference value for that variable (in the case of age group the reference value is 65-69 years of age). Reporting APEs has the advantage that effects can be compared across groups, across samples, and across models (Mood, 2010). APEs were calculated with standard errors and 95% confidence intervals using the R package `margins` v.0.3.23 (Leeper, Arnold and Arel-Bundock, 2017).

1.6.2 Research question 2

Likely to be the same as above so may not need a separate section but some additional description above.

1.7 Timeline

Figure 1.2: Timeline of Thesis project

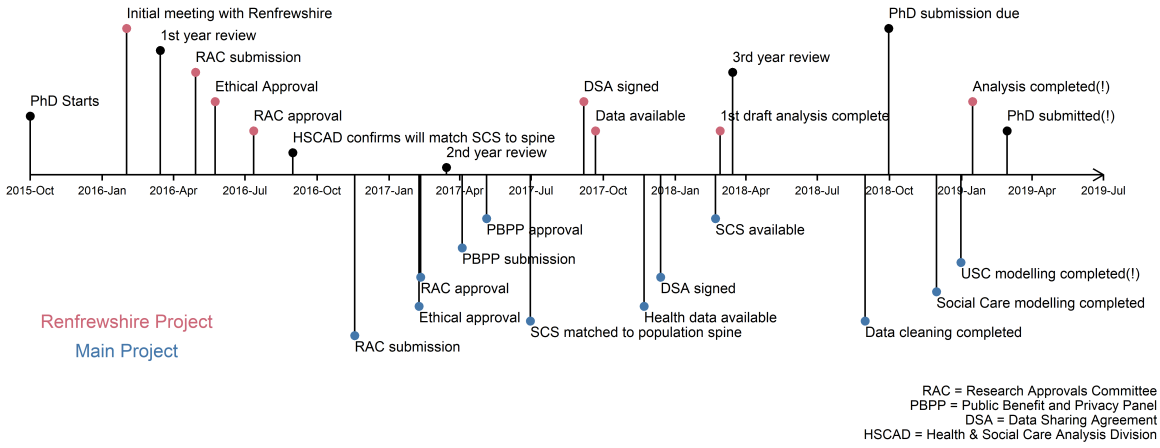


Figure 1.2 depicts major milestones during the thesis project, including a separate analysis conducted with data from Renfrewshire council. The journey through the approvals and analysis process for that project is reported in chapter **??**, therefore this description focusses on the main analysis project. Renfrewshire information is depicted to provide context to the time frame of the whole project, in particular the availability of that data arriving only shortly before the availability of data for the main project.

Part of the first year of the thesis was spent scoping potential sources of administrative social care data and appraising their usefulness for research purposes. It was not until near the beginning of the thesis second year, August 2016, that HSCAD at the Scottish Government confirmed it would make the Social Care Survey available for linkage. This process was not completed until the end of June 2017 (section 1.4).

Approvals process through the UBDC RAC was governed by its bi-annual submission process - with approval for the main project being requested in Autumn 2016. Approval from this committee was necessary before applying to the PBPP which gave its approval for the main project in May 2017, subject to completion of a DSA between relevant parties. This was finally signed (Appendix D) in November 2017. Some health data from ISD was made available for analysis slightly prior to this date, and shortly after data from the

Renfrewshire project had been made available. Social care data was finally transferred, enabling full cleaning and analysis for the main project to commence, in February 2018.

References

Ai, C. and Norton, E. C. (2003) 'Interaction terms in logit and probit models', *Economics letters*, 80(1), pp. 123–129. Available at: https://pdfs.semanticscholar.org/6285/8e64d9a337504d72cb862c4cc1e7fd27a7a0.pdf.

Alvarez-Madrazo, S. *et al.* (2016) 'Data resource profile: The scottish national prescribing information system (pis)', *International journal of epidemiology*, p. dyw060. Available at: http://ije.oxfordjournals.org/content/early/2016/05/09/ije.dyw060.extract.

Atherton, I. M. *et al.* (2015) 'Barriers and solutions to linking and using health and social care data in scotland', *British Journal of Social Work*, 45(5), pp. 1614–1622. doi: 10.1093/bjsw/bcv047.

Bell, J. *et al.* (2016) *Legal issues for adrn users*. Report. Administrative Data Research Network. Available at: https://adrn.ac.uk/media/174198/legal_guide_final.pdf.

Doidge, J. and Harron, K. (2018) 'Demystifying probabilistic linkage: Commin myths and misconceptions', *International Journal of Population Data Science*, 3(1), pp. 1–8. Available at: https://ijpds.org/article/view/410/384.

Dowle, M. *et al.* (2018) *Data.table: Extension of 'data.frame'*. Report. Available at: https://CRAN.R-project.org/package=data.table.

Fleming, M., Kirby, B. and Penny, K. I. (2012) 'Record linkage in scotland and its applications to health research', *Journal of clinical nursing*, 21(19pt20), pp. 2711–2721. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2702.2011.04021.x/full.

Grolemund, G. and Wickham, H. (2017) *Lubridate:Dates and times made easy with lubridate. r package version 1.6.0.* Report. Available at: https://CRAN.R-project.org/package=lubridate.

Guthrie, B. *et al.* (2015) 'The rising tide of polypharmacy and drug-drug interactions: Population database analysis 1995–2010', *BMC medicine*, 13(1), p. 74. Available at: http://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-015-0322-7#MOESM1.

Harron, K. (2016) *An introduction to data linkage*. Report. University of Essex.

Harron, K. *et al.* (2017) 'Challenges in administrative data linkage for research',

Big Data and Society, 4(2). Available at: http://journals.sagepub.com/eprint/wghywxV6WvbGZtRzjE5R/full#articleShareContainer.

Hashimoto, R. E. *et al.* (2014) 'Administrative database studies: Goldmine or goose chase?', *Evidence-based spine-care journal*, 5(2), p. 74. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4174180/pdf/10-1055-s-0034-1390027.pdf.

Henry, L. and Wickham, H. (2017) *Purrr:Functional programming tools. r package version 0.2.4.* Report. Available at: https://CRAN.R-project.org/package=purrr.

ISD (2010a) *EDRIS frequently asked questions*. Report. Available at: http://www.isdscotland.org/Products-and-Services/eDRIS/FAQ-eDRIS/index.asp#d4.

ISD (2010b) *Information services division*. Report. Available at: http://www.isdscotland.org/.

ISD (2010c) *Use of the nss national safe haven*. Report. Available at: http://www.isdscotland.org/Products-and-services/Edris/Use-of-the-National-Safe-Haven/.

ISD (2017) *Urgent care data mart (ucd) - background paper*. Report. Available at: http://www.isdscotland.org/Health-Topics/Emergency-Care/Patient-Pathways/UrgentCareDataMartBackgroundPaper_20171002.pdf.

Leeper, T., Arnold, J. and Arel-Bundock, V. (2017) *Margins: Marginal effects for model objects. r package v0.3.23*. Report. Available at: https://cran.r-project.org/web/packages/margins/index.html.

Louviere, J. J., Hensher, D. A. and Swait, J. D. (2000) *Stated choice methods: Analysis and applications*. Cambridge: Cambridge university press.

Lowthian, P. and Ritchie, F. (2017) *Ensuring the confidentiality of statistical outputs from the adrn*. Report. Available at: https://www.adrn.ac.uk/media/174254/sdc_guide_final.pdf.

Mazzali, C. and Duca, P. (2015) 'Use of administrative data in healthcare research', *Internal and emergency medicine*, 10(4), pp. 517–524. Available at: https://link.springer.com/article/10.1007/s11739-015-1213-9.

McFadden, D. (1974) 'Conditional logit analysis of qualitative choice behavior', in P., Z. (ed.) *Frontiers in econometrics*. London: Academic Press. Available at: https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf.

McFadden, D. (1977) *Quantitative methods for analyzing travel behavior of individuals: Some recent developments*. Yale: Cowles foundation for research in economics. Available at: https://core.ac.uk/download/pdf/6448852.pdf.

Mood, C. (2010) 'Logistic regression: Why we cannot do what we think we can do, and what we can do about it', *European sociological review*, 26(1), pp. 67–82. Available at:

https://academic.oup.com/esr/article/26/1/67/540767.

Mustillo, S. A., Lizardo, O. A. and McVeigh, R. M. (2018) 'Editors' comment: A few guidelines for quantitative submissions', *American Sociological Review*. Available at: https://journals.sagepub.com/doi/full/10.1177/0003122418806282?casa_token=quE2fKGhcJAAAAAA:2JQiLy31Ni2tIQFler0c268rjAZgNUWzah-adAiZmn1rmqmxMPhRpMrR4nzQnFI

NRS (2017) *About us*. Report. Available at: https://www.nrscotland.gov.uk/about-us.

NRS (2018a) *NHS central register (nhscr)*. Report. Available at: https://www.nrscotland.gov.uk/statistics-and-data/nhs-central-register.

NRS (2018b) *Statutory registers of deaths*. Report. Available at: https://www.nrscotland.gov.uk/research/guides/statutory-registers/deaths.

NSS (n.d.) *How nss works*. Report. Available at: https://nhsnss.org/how-nss-works/.

Pavis, S. and Morris, A. (2015) 'Unleashing the power of administrative health data: The scottish model', *Public Health Research and Practice*, 25(4), p. e2541541. Available at: http://www.phrp.com.au/issues/september-2015-volume-25-issue-4/unleashing-the-power-of-administrative-health-data-the-scottish-model/.

R-Core-Team (2017) *R: A language and environment for statistical computing. r foundation for statistical computing*. Report. Available at: http://www.R-project.org.

RStudio-team (2016) *Integrated development for r. rstudio, inc., boston, ma. version 1.0.143*. Report. Available at: http://www.rstudio.com/.

Scottish-Government (2012) *Guiding principles for data linkage*. Report.

Scottish-Government (2016a) *2016-17 settlement. grant aided expenditure green book*. Report. Available at: http://www.gov.scot/Resource/0049/00499184.pdf.

Scottish-Government (2016b) *Scottish index of multiple deprivation 2016*. Report. Available at: http://simd.scot/2016/#/simd2016/BTTTFTT/9/-4.0011/55.9001/.

Scottish-Government (2016c) *Social care services, scotland, 2016*. Report. Available at: https://www.gov.uk/government/statistics/social-care-services-scotland-2016.

Scottish-Government (2017a) *Inpatient census, 2017*. Report. Available at: http://www.gov.scot/Resource/0052/00524621.pdf.

Scottish-Government (2017b) *Social care survey*. Report. Available at: http://www.gov.scot/Topics/Statistics/Browse/Health/Data/HomeCare.

Scottish-Government (n.d.) *Public benefit and privacy panel for health and social care*. Report. Available at: http://www.informationgovernance.scot.nhs.uk/pbpphsc/.

SILC (2017a) *Data linkage safeguards*. Report. Available at: http://www.

datalinkagescotland.co.uk/data-linkage-safeguards.

SILC (2017b) *EDRIS*. Report. Available at: http://www.datalinkagescotland.co.uk/edris.

SILC (2017c) *Home*. Report. Available at: http://www.datalinkagescotland.co.uk/.

UBDC (2017a) *Controlled data service guide for researchers [last updated 01.02.2017]*. Report. Available at: http://ubdc.ac.uk/media/1445/ubdc-controlled-data-services-guide-for-researchers-v5 pdf.

UBDC (2017b) *Overview [accessed 14.11.2017]*. Report. Available at: http://ubdc.ac.uk/about/overview/.

UBDC (2017c) *Resarch approvals committee*. Report. Available at: http://ubdc.ac.uk/about/overview/research-approvals-committee/.

Walesby, K., Harrison, J. and Russ, T. (2017) 'What bid data could achieve in scotland', *Journal of the Royal College of Physicians of Edinburgh*, 47(2), pp. 114–119. doi: 10.4997/JrCPe.2017.201.

Walraven, C. van and Austin, P. (2012) 'Administrative database research has unique characteristics that can risk biased results', *Journal of clinical epidemiology*, 65(2), pp. 126–131. Available at: https://www.sciencedirect.com/science/article/pii/S0895435611002484.

Wickham, H. (2014) 'Tidy data', *Journal of Statistical Software*, 59(10), pp. 1–23. Available at: https://www.jstatsoft.org/article/view/v059i10.

Wickham, H. (2015) 'Stringr: Simple, consistent wrappers for common string operations', *R package version*, 1(0).

Wickham, H. (2017) *Forcats:Tools for working with categorical variables (factors). r package version 0.2.0.* Report. Available at: https://CRAN.R-project.ork/package=forcats.

Wickham, H. and Francois, R. (2017) *Dplyr: A grammar of data manipulation. r package version 0.7.4.* Report. Available at: https://CRAN.R-project.org/package=dplyr.

Wickham, H. and Grolemund, G. (2017) *R for data science: Import, tidy, transform, visualize, and model data*. 1st ed (2nd release 2016-12-22). Sebastopol, CA: O'Reilly Media, Inc. Available at: https://r4ds.had.co.nz/.

Wickham, H. and Henry, L. (2017) *Tidyr:Easily tidy data with 'spread()' and 'gather()' functions. r package version 0.7.2.* Report. Available at: https://CRAN.R-project.org/package=tidyr.

Williams, R. (2012) 'Using the margins command to estimate and interpret adjusted predictions and marginal effects', *Stata Journal*, 12(2), p. 308. Available at: https://ideas.repec.org/a/tsj/stataj/v12y2012i2p308-331.html.

Witham, M. D. *et al.* (2015) 'Construction of a linked health and social care database

resource–lessons on process, content and culture', *Inform Health Soc Care*, 40(3), pp. 229–39. doi: 10.3109/17538157.2014.892491.

Wood, R. *et al.* (2013) 'Novel cross-sectoral linkage of routine health and education data at an all-scotland level: A feasibility study', *The Lancet*, 382, p. S10. Available at: http://www.sciencedirect.com/science/article/pii/S0140673613624356.

Zeileis, A. *et al.* (2018) *Zoo: S3 infrastructure for regular and irregular time series (z's ordered observations)*. Report. Available at: https://CRAN.R-project.org/package=zoo.