Dadansoddi ar gyfer Polisi

YMCHWIL GYMDEITHASOL Y LLYWODRAETH

GSR

GOVERNMENT SOCIAL RESEARCH

Analysis for Policy

Llywodraeth Cymru
Welsh Government

www.cymru.gov.uk

Ymchwil gymdeithasol

Social research

Number: 76/2013

Data Linking Demonstration Project - Journey Mapping for Patients with Multiple Chronic Health Conditions

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

Martin Heaven, Health Information Research Unit Swansea University / Data Linkage Fellow, Welsh Government

Sarah Lowe, Knowledge and Analytical Services, Welsh Government

Views expressed in this report are those of the researchers and not necessarily those of the Welsh Government

For further information please contact:

Name: Sarah Lowe

Department: Knowledge and Analytical Services

Welsh Government

Cathays Park
Cardiff

CF10 3NQ

Tel: 029 2082 6229

Email: sarah.lowe@wales.gsi.gov.uk

Welsh Government Social Research, 2013

ISBN 978-1-4734-0673-5

Table of contents

Glossary of acronyms

A&E	Accident and Emergency (Data Set)
ALF	Anonymised Linking Field
CHS	Child Health (System Data Set)
EASHR	European Age Standardised Hospitalisation Rate
EASMR	European age standardised Mortality Rate
EDDS	Emergency Department Data Set
FEAS	First Emergency Admission for Stroke (local to this document)
HES	Hospital Episode Statistics
HIRU	Health Information Research Unit
IGRP	Information Governance Review Panel
LSOA	Lower Super Output Area
MCHC	Multiple Chronic Health Conditions
NWIS	NHS Wales Information Service
NHSAR	NHS Administrative Register, (now superseded by WDS)
ONS	Office for National Statistics
PAS	Patient Administration System
PEDW	Patient Episode Database for Wales
PHW	Public Health Wales
RALF	Residential Anonymised Linking Field
SAIL	Secure Anonymised Information Linkage
SQL	Structured Query Language
WDS	Welsh Demographics Service (GP registration history database)
WG	Welsh Government
WIMD	Welsh Index of Multiple Deprivation

1 Introduction

The aim and objectives of the demonstration project

1.1 This project is being delivered as part of the Welsh Government Programme to Maximise the Use of Existing Data. The programme aims to demonstrate the unique contribution data linking can make to the evidence base. The suite of three data linking demonstration projects has examined the anonymised data linkage process from acquiring additional data to carrying out analysis on new data sets created by linking existing administrative data. The projects are intended to stimulate engagement of appropriate WG officials with regard to information governance and practical issues around acquiring, processing and analysing new linked data sets. The projects were designed to be small in scale and exploratory in nature. These constraints are reflected in their relatively limited scope and in both the practical and analytical decisions made throughout.

1.2 This Project did not require the inclusion of any new datasets into the SAIL databank at Swansea University, but demonstrated the processes involved in linking numerous existing health services data flows to analyse the ‘patient pathways’ for groups of patients with similar conditions.

1.3 The number of hospital spells in Wales has shown a general increase over recent years[1]. Since many health issues are age related, this increase is partly due to the increasing number of older people in the population; however, it may also be explained by advances in medical science (it is now possible to treat conditions differently and possibly to treat people who in the past would be considered too frail) and more modern treatment regimens[2],[3] which, for example, allow the patient to be hospitalised for a shorter period but may require more visits during the course of the treatment. Any project providing insights about effective treatment regimens or patient specific pathways will be useful in designing a more efficient and effective health service for the future.

[1] http://wales.gov.uk/topics/statistics/headlines/health2013/health-statistics-wales-2013/?lang=en

[2] BMJ VOLUME 307 1 1 DECEMBER1993, Minimally Invasive Surgery : Implications for hospitals, health workers, and patients H David Banta. Accessed at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1679562/pdf/bmj00051-0045.pdf

[3] Use of PET scanning in diagnosis: http://www.sarcoidosisonlinesites.com/cardiac_sarcoidosis.pdf

However, the reader should bear in mind that the Project was designed to demonstrate the usefulness of linked data in allowing patient pathways to be developed; findings that relate to health service use are therefore presented as steps along the way to achieving that objective, not to serve as a report on the health of the population of Wales. All of the indicators used are indicative and for demonstration purposes only. In any future research project building on these methods, a multi-disciplinary team of experts, including policy makers, would contribute to the process of identifying appropriate indicators.

1.4 The focus of this project was to explore the extent to which the analysis of linked datasets can provide evidence to help policymakers understand the complexity of managing patients with multiple health conditions. With this objective in mind, the Project used various methods to try to identify groups of patients with similar 'pathways', including the experimental use of 'cluster' or 'segmentation' analysis. These methods are more commonly used by marketing analysts to 'segment' the population into groups of 'customers' with similar needs or interests in order to target their communications activities more effectively. In this case, it allowed groups of patients with similar pathways to be identified in order to examine their previous interactions with health services.

1.5 Within the limited scope of a demonstration project, it was not possible to examine every possible combination of chronic health conditions. The Project demonstrated that the use of linked administrative health data sets is even more complex and time-consuming than had been imagined. A significant amount of initial analysis was done both to define possible chronic conditions and to identify the possible indicators of health service use. The initial proposal was to examine combinations of the following conditions, each of which was a particular focus for policy:

- stroke / cardiovascular disease;
- diabetes;
- respiratory disease (asthma and chronic obstructive pulmonary disease);
- coronary heart disease; and
- arthritis.

1.6 After the initial analysis demonstrated how time-consuming it was to deliver analysis for just a single condition, it was agreed to focus on two groups, those with strokes and, within that group, those with diabetes. This Project also used a relatively limited set of health service use indicators. However, the initial work completed to define additional chronic conditions and indicators of health service use has been documented so that future analysts can make use of them.

1.7 The process, issues, problems and limitations encountered are documented in the following chapters. Chapter 2 describes the process by which linked administrative data are made available for research purposes. Chapter 3 describes the datasets used in this report. Chapter 4 describes the methodologies applied to the datasets to identify a set of stroke patients and summarise aspects of their health service utilisation by creation of a series of indicators. Chapter 5 summarises the characteristics of the set of stroke patients. Chapter 6 uses a small subgroup of the stroke patients (younger men) to further explore the available data and the usefulness of the indicators. Chapter 7 explores the effects of patients having a single complicating condition or co-morbidity[4] on outcomes for stroke patients. Chapter 8 describes the results of a cluster or segmentation analysis of stroke patients. Chapter 9 summarises the overall findings and suggests how the Project might be further developed.

1.8 The challenges that emerged during the demonstration project will be explored further in a lessons learned report spanning all three data linking demonstration projects, publication of which is to follow. Lessons will be fed back into SAIL processes under development to make the research process more user friendly.

[4] In medicine, co-morbidity is either the simultaneous presence of one or more disorders (or diseases) in addition to a primary disease or disorder, or the effect of such additional disorders or diseases.

2 Making health data available for research

The development of SAIL

2.1 Data collected by NHS organisations is usually recorded as part of the face to face interaction with patients. It is necessary to hold some patient identifiers as part of these data sets, for example full name, address, postcode, date of birth and gender are all used in establishing identity. NHS numbers are not memorable enough to be practical in a face to face situation, so are attached to records by a look-up system that utilises the kinds of identifiable information listed above. As a consequence, different organisations within the NHS hold their own identifiable datasets in very secure and often completely separate environments. Information Governance law prohibits the use of identifiable data in this form for research outside the NHS and NHS professionals have signed up to a code of confidentiality surrounding their interaction with patients. These restrictions would potentially leave large amounts of useful research data unavailable for research.

2.2 However, research that focuses on the characteristics of - and what happens to - large numbers of patients does not need to identify individuals as long as information about the same person that is stored in the different systems can be reliably linked. Some mechanism for safely processing identifiable data from separate NHS systems, without breaching confidentiality, into anonymised linked data for research was therefore required.

2.3 To address this issue, the Welsh Government (WG) funded the creation and development of the Health Information Research Unit (HIRU) at Swansea University, from 2006. One aim of this unit was to develop a means by which routinely collected health data from many different sources could be utilised in a linked way, but held in such a way that the data is completely anonymous and the identity of individuals cannot be revealed. The process developed was called Secure Anonymised Information Linkage (SAIL); it led to the collation of a very large collection of anonymised health datasets stored in Swansea University and made available only to approved researchers and purely for research purposes. The process can also be applied to other datasets and data for topics such as education and housing as well as survey and other data is now being added to the growing collection of data. The name of ‘SAIL’ has become synonymous with both the large data repository and the HIRU. ‘SAIL’ will be used to refer to the databank at Swansea University throughout the remainder of this document.

2.4 In the data anonymisation process, the NHS Wales Information Service (NWIS) acts as a trusted third party indexing organisation. They are provided with only the identifiable components of each dataset. The identifiable data is provided either at individual person level or address level. NWIS use the Welsh Demographic Service (WDS) data as the ’population spine’ or ‘template’ for its anonymisation process. The WDS is a database of everyone registered with a GP in Wales from 1994 to the present day. It includes an anonymised residential address history – an index of numbers, one for each household in Wales, known as the Residential Anonymised Linking Field (RALF). Individual people who have been registered with a GP in Wales, past and present, are represented in the WDS data as another index of unique numbers, known as the Anonymised Linking Field (ALF). In this way, it is possible to associate ALFs with RALFs, that is: people to homes.

Information Governance Issues

2.5 SAIL follows the data protection guidance provided by the Data Commissioner’s Office, and operates within the Swansea University Data Protection policy which is in line with all the relevant UK laws. The anonymous nature of data held in SAIL is such that it is not governed by the Data Protection Act, and it has been agreed by the National Research Ethics Service that research carried out within SAIL does not require ethical review. However all research carried out within SAIL is still managed through a rigorous control structure to ensure that confidentiality is maintained and potentially disclosive outputs are not produced.

2.6 One of the controls in place is a requirement for all proposals involving the analysis of linked data within SAIL to obtain approval from the Information Governance Review Panel (IGRP). IGRP is a panel of independent specialists in informatics governance and lay members that oversee all research taking place within SAIL. Current membership (June 2013) is listed in Appendix 2. An

IGRP application contains an outline of the research rationale for creating the links, any new datasets that would be accessed, and precisely what variables would be required from the linked datasets. Researchers must indicate in the application that they have considered the handling of sensitive data in the research design. Although the data sets are all totally anonymised in SAIL, the selection of a really specific sub-group based on age and gender at small area (LSOA) level, looking at a specific condition could return small numbers. Small numbers in a published output could be put together with other local knowledge to establish who the statistic refers to. Researchers are given access to the data at the level of detail necessary in order to complete their analysis, but need to ensure that nothing potentially identifiable is revealed in their reporting. IGRP applications must indicate how the analyst proposes to deal with small numbers (e.g. through grouping and aggregation of cases).

2.7 Although the data sets are all totally anonymised in SAIL, the selection of a really specific sub-group based on age and gender at small area (LSOA) level, looking at a specific condition could return small numbers. Small numbers in a published output could be put together with other local knowledge to establish who the statistic refers to.

2.8 Researchers are given access to the data at the most detailed level in order to complete their analysis, but need to ensure that nothing potentially identifiable is revealed in their reporting. IGRP applications must indicate how the analyst proposes to deal with small numbers (e.g. through grouping and aggregation of cases).

2.9 The IGRP application for this Project, (Appendix 2) was amended prior to submission for IGRP consideration, following feedback from the SAIL management team, who suggested that the original version was not specific enough. The potential linked dataset originally described was huge. Following discussions with the WG team, some data limiting criteria were written into the application, which was then successful without any concerns being raised.

2.10 This may be in part due to the research fellow being very familiar with the application process and having discussed the data maximisation demonstration projects at a prior meeting of the IGRP panel. The process still took eight weeks to complete, mainly due to waiting for individual reviewers to respond. The target time to from submission to approval for a project is 4 weeks, but the IGRP panel members perform this role in their own time, which for some of them is a scarce commodity. Researchers should be aware that gaining IGRP approval can be a time-consuming, iterative process requiring adjustments to their research proposal. Lessons learned during the implementation of this and the other demonstration projects are presented in the report “Lessons learned from the Data linking fellowship 2011-13”, and will also feed back into the developments in place to streamline the data access process. As noted above, a lessons learned report will follow.

Transferring data into SAIL

2.11 The normal standard practice for transferring data is to utilise a secure electronic data transfer facility. For NHS organisations transfers into NWIS use such a system based on the Digital All Wales Network (DAWN). For non-NHS data providers, a secure internet based facility is in place, and for the transfer of data into SAIL a separate but similar Internet based facility is available[5]. Data sharing agreements are in place for all the datasets used for this project, and the data are regularly flowing into SAIL by this means.

2.12 The data are transferred using what is known as the “split file process”. This is a common procedure for safeguarding respondent privacy during data linking Identifying information i.e. name, address, date of birth, NI number etc. is separated from all other analytical data, whether medical, social, financial, attitudinal etc. in each source to be linked. For each source, this creates two files, the first containing an index plus the identifiable information and the second containing an index plus the analytical data. For each dataset the identifying information is sent to a trusted third party who creates an anonymous linking field. Once the linking field has been created the identifying information is destroyed leaving only the linking field and the index. The index allows the anonymous linking field to be reattached to the analytical data. The analytical data can then be linked to other anonymised data sets without using

[5] See secure data transportation in “The SAIL Databank: building a national architecture for e-health research and evaluation” http://www.biomedcentral.com/1472-6963/9/157 .

any identifying variables. This is the procedure used by SAIL, among others, to create linked datasets.

Gaining access to the data

2.13 A database “view” is a structured ‘image’ of information stored in the database, including only a subset of the complete dataset. A view can include data from more than one database, and can be restricted to include specific rows and columns. In this way, the database administrator can very closely control the data each researcher can work with. There is, in addition, no way that a researcher can alter the underlying data table providing the “view”.

2.14 The “view”, tailored specifically to meet the requirements of the researcher’s project, is loaded into the SAIL databank by the SAIL technical team. The SAIL technical team provide the hardware and database management support for research and are not data analysts. Separating the data management and research analyst functions prevents the need for technical team members to understand the data and for researchers to access underlying data tables or any intermediate stage data. Access to views is controlled and restricted to authorised approved researchers. For the Multiple Chronic Health Conditions Project, access was originally restricted to the Author, who is a senior research analyst working in the SAIL databank. Later, access was sought for named WG staff for Quality Assurance purposes.

2.15 The database views are made available through a secure remote access system, the SAIL Gateway, which can be accessed securely over the internet, using a system where authorised researchers are able to log on to a dedicated computer through a password protected browser. Outputs are ‘locked down’, so that nothing can be copied and pasted out of the gateway, saved to a port or drive on the remote computer, or sent to a printer.

2.16 All analysts who are provided with a SAIL gateway account are given access only after both they and their line manager have signed a detailed agreement outlining the researcher’s responsibilities and the agreed usage that can be made of that account. The agreement clearly places the responsibility with the researcher to ensure that no individual could be potentially identifiable from the research outputs. However, in addition, all potential outputs are scrutinised by a SAIL administrator to ensure potentially disclosive information does not leave the secure gateway.

2.17 The researcher is required to carry out the analysis within the gateway, in which suitable database, statistical, spreadsheet, word processing, mapping and presentation software are available. The only outputs allowed are summarised or aggregate results. Proposed outputs are processed through the ‘request data out’ link within the gateway. This is the stage at which outputs are scrutinised by a senior research analyst in SAIL, checking for potential disclosure issues such as small numbers. The ‘data out’ process does not check that the analysis has been performed correctly and that results are correct, it merely scrutinises outputs for potentially disclosive situations.

2.18 It is not possible to put a process in place that would stop researchers taking photographs of a computer screen, for example, or simply writing down results and not following the ‘data out’ procedure. Given this, the researcher must be trusted to adhere to the terms of the SAIL access agreement. However, when signing the access agreement, researchers and their line managers are agreeing to abide by the statement of procedures in the National Statistics Code of Practice: Protocol on Data Access and Confidentiality, in both letter and spirit, to the maximum extent that they apply. Breaches of these rules would result in penalties and legal action. As part of the creation of the UK Administrative Data Research Network, only ‘accredited researchers’ will be allowed to apply to access databases like SAIL. Abuse of the privileges of data access would result in removal from such a register, effectively ending the perpetrator’s research career.

2.19 Access controls are managed by SAIL administrators utilising database views. A database “view” is a structured ‘image’ of information stored in the database, usually including only a subset of the complete dataset. A view can include data from more than one database table, and can be restricted to include specific rows and columns. Researchers working in SAIL are given access to views of database tables. In this way, what they are able to see can be heavily controlled by the database administrator, and there is no way that a researcher can alter the underlying data table providing the “view”.

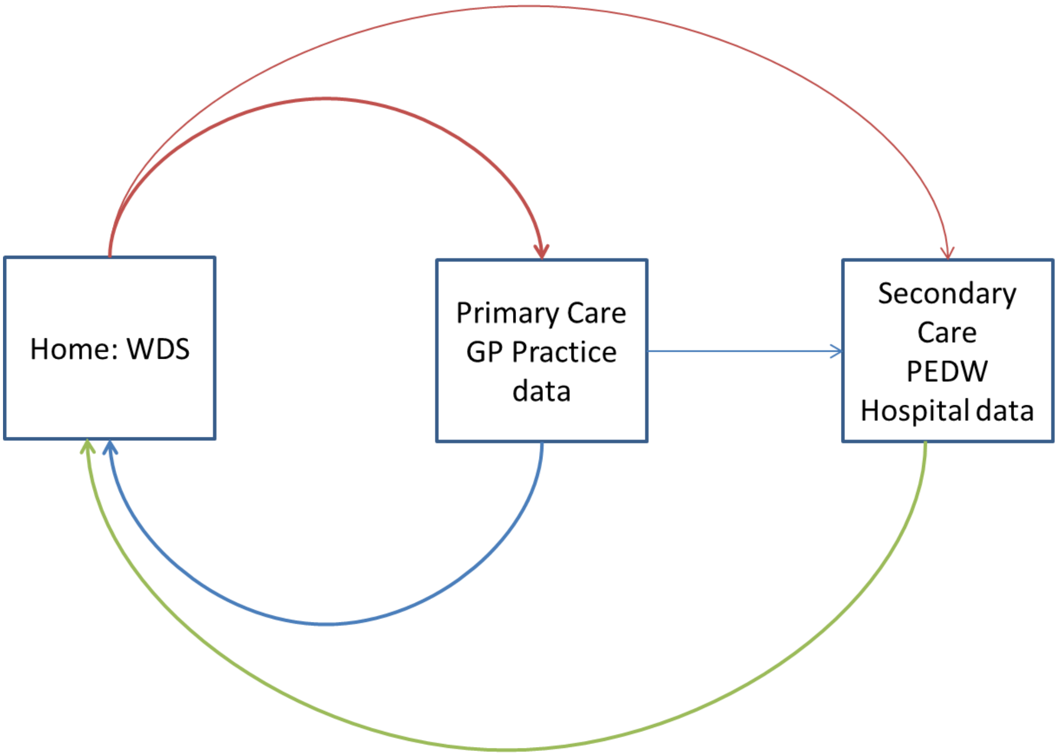
2.20 Detailed descriptions of the datasets available in SAIL that have been accessed for this project are provided in Appendix 3, with lists of available fields in Appendix 4. Issues related to the content of these fields are discussed in relevant sections of this document.

3 The Datasets used in this Report

3.1 A key objective of this Project was to look for patterns in existing, linked health data that indicate similarities in patients' 'care pathways', and to summarise these into groups of patients with similar experiences. For patients with multiple chronic conditions, the 'care pathway' can be visualised as a continuous cycle of events as illustrated in Figure 3.1 below. Firstly there is a continuous cycle of interaction between the individual and primary care, representing the detection and management of conditions by the patient supported by a primary care practitioner. The frequency of the primary care cycle varies according to many different factors and ranges from virtually no contact with the GP practice to almost daily interactions. When conditions require, some patients will also require treatment in secondary care. This may be a planned event (for specific diagnostic procedures or planned operations) or be an unscheduled or emergency event.

3.2 Hospitalisation ranges from a single admission to a series of admissions interspersed with support activity from primary and community care. This Project seeks to find patterns in these cycles that can be used to make some sense of these complicated pathways, and establish if they vary for different patient groups. Some examples of the variables available to achieve this task are described in detail in this Chapter. At point of writing, sufficient data for community care was not available in SAIL but the project has led to the commissioning of a WG-funded Project to Improve Linked Data for Social Care; the Project therefore focuses on information on primary and secondary care.

Figure 3.1 Visualisation of the Project 'care pathway'


Home: WDS
Primary Care GP Practice data
Secondary Care PEDW Hospital data

3.3 Figure 3.1 is a very simplified illustration of the patient pathway, not depicting Accident and Emergency department and outpatient department visits, care provided by Social Services, or the end point of the cycle, i.e. death. The developing methodology does not currently make use of A&E and outpatient data but could be expanded to include other such datasets. Suggested further work is listed in Appendix 5.

3.4 Figure 3.1 lists the source of the information used for each part of the care pathway. These are described in more detail below, including a more detailed description of some of the variables chosen for analysis in this Project and pointing out some of the known issues with the data.

Patient Registration Data – Welsh Demographic Service

3.5 The Project defined a set of patients based on them having similar health conditions and then established, for each patient, a history of their interactions with health services. In order to establish the patient's age and gender, changes in registration between different GP practices, and the total period of time they have been a patient in the NHS in Wales, the Welsh Demographic Service data has been used.

3.6 The WDS database exists because almost all residents of Wales are registered with one of around 500 GP practices distributed throughout the country. Payments to these GP practices are based on a complex funding formula that adjusts payment according to the weighted needs of the population. So, for example, a practice with a large elderly population gets a higher payment than a practice with a relatively younger population[6]. A detailed reporting system is therefore required to track exactly which people are registered with each practice and the WDS serves this purpose. The picture is constantly changing as new babies are born and registered, other people's lives end, and there are migrations both within Wales and in and out of Wales from and to other parts of the world.

3.7 The dataset includes administrative information about all individuals resident in Wales who have used NHS services. It replaced the NHS Wales Administrative Register (NHSAR) in 2009. The dataset contains the full registration history of the population of Wales since 1990, including house moves and changes of registration to different GP practices. This is the core data that is used in linking datasets together in SAIL. Each person's week of birth is recorded and, when known, a date of death[7]. Dates of birth are changed to week of birth as part of the anonymisation process, and all addresses are anonymised so that it is not possible to locate them geographically more precisely than at the Lower Super Output Area (LSOA) level – LSOAs are a patchwork of small areas covering the UK, each one of which contains on average 1600 people. A typical patient registration history is described in the box below.

[6] http://bma.org.uk/-/media/Files/Word%20files/News%20views%20analysis/pressbriefing_GPs.doc .

[7] While Date of Birth is restricted to 'Week of Birth' in SAIL, Date of Death is not restricted to 'Week of Death'. The problem is that if we restrict DOD, there are so few deaths in women of child-bearing age that they would still potentially be identifiable. However, if we restrict further, most mortality analyses (e.g. 30 day) become very difficult to implement. The solution lies in the encrypting of LSOA plus the restricted access provided to the data and the Data Access Agreement researchers sign, in which they promise not to attempt any such analyses and there are penalties attached for breaches. Most importantly, external researchers do not see real LSOA codes, only encrypted ones.

Mr Smith's GP registration history (fictitious but based on real WDS data)

Mr Smith was born on or around 28.11.1927 but we don't know where. He was registered with a GP practice in Wales on 29.06.1974 at the age of 46 but the first address registration date recorded for him was 03.03.1987, thirteen years later when he was 59. The next house move recorded (chronologically) says he moved house on 05.01.1993 but another record indicates that he moved back into the same house on the same date. He moved house again on 06.12.2002 and this time it was to a different location. He lived there until 05.01.2012, which is the last known address for him.

Mr Smith died on 13.08.2012. This was also the date recorded as the end of his last registration with a GP. During his life in Wales he was registered with one GP practice, but there are three sets of end dates for that registration, followed by three sets of registration dates with the same practice, none of which coincide with his house moves. These may be due to a change of GP within the practice, but the anonymised data does not hold this detail.

Note that there are two spells of time when Mr Smith was registered with a GP in Wales but for which we do not have his address. The first was for 13 years at the beginning of his registration in Wales and the last was a few months at the end of his life, when perhaps he was living in some kind of care facility.

3.8 The WDS data is deceptively simple, since it contains relatively few variables – so, for example, the beginning and end of each GP practice registration period is defined by a pair of dates and periods of residence at particular (anonymised) addresses are defined by a pair of dates. In order to establish comparable event rates for the project, it is necessary to work out how long each individual has been contributing to the datasets.

3.9 However, the WDS in fact contains thousands of records that indicate that an individual left a property on a specific date, only to move into that property again either on the same date or a few days later. This problem may be due to software and hardware changes to the data collection systems, such as when upgrades occur at a GP practice. These 'false leaving dates' create difficulty when attempting to select people with continuous residence over a specific period of time. The simplest algorithm to establish if someone was living in the same place during the Project time period, would be **“Select all people who moved in on or before 01.01.2000 and who did not move out until on or after 31.12.2011.”** However if a person has two records that effectively split a continuous period of residence into two components, as in Table 3.1 below, the algorithm would not select either record, resulting in the individual being wrongly excluded from the analysis. For this Project we are not selecting a group of people based on their residence at a particular address but are analysing health service use over time, despite changes in residence and GP registration, so we do need to know the total period of time they were registered for healthcare with the NHS in Wales. This is achieved by summing the periods of time between ‘from dates’ and ‘to dates’.

3.10 Just as we have records of ‘from dates’ and ‘to dates’ for patients’ (anonymised) home addresses, we have ‘from dates’ and ‘to dates’ for periods of registration with a GP practice. These are useful when comparing repeating occurrences over time, e.g. when comparing prescribing rates over time, it is essential to account for the total period of time when each patient was available (registered) to be prescribed to. These records also suffer from the ‘false leaving dates’ problem and have to be carefully processed to eliminate its effects.

Table 3.1: Sample moving-in and moving-out dates

ALF	RALF	Move-in Date	Move-out Date	Notes
12412341234	789798748	16.04.1999	04.05.2006	Same person ID and same address ID for both records.
12412341234	789798748	04.05.2006	31.03.2013	

Source: Fictitious data based on SAIL records

General Practice Data – The Primary Care GP Dataset

3.11 SAIL contains data from around 42% of the GP practices in Wales, which includes detailed data on primary care activity for around 47% of the population of Wales.

3.12 GP Practice Data is coded using a hierarchical scheme of codes called Read codes, of which there are two versions currently in use, Version 2 and Version

3. Most of the GP practices providing data into SAIL are using Version 2, so the analyses developed for this report are based on this version. However these analyses should eventually be adapted to include Version 3 codes. It is not clear at the time of writing how much the non-inclusion of version 3 codes is having an effect on analyses. Other SAIL projects are currently assessing this issue.

3.13 The Read hierarchy comprises eight sections:

- History, examination and observations.
- Investigations.
- Operations and Procedures.
- Disorders.
- Administration.
- Drug and Appliance Products.
- ICD10 Disease codes.
- OPCS4 Operative procedure codes.

There is considerable overlap in these sections, which means that there are often multiple places in the coding structure where relevant detail can be recorded. This makes the selection of a set of suitable codes to find particular characteristics of patients a time consuming and complex process, sometimes with imprecise results.

3.14 Data may be recorded as the result of patients attending the practice, where they may be seen by a nurse, a doctor, specialist staff and administrative staff or because a repeat prescription has been requested. Results of tests ordered or happening in other parts of the health care system are fed into the system, e.g. details from hospital discharge letters. Over time, a great many records are generated about each patient, each with coded ‘event’ and associated date, and some with an associated event value e.g. a code indicating a blood pressure check was made is sometimes accompanied by the actual BP reading.

3.15 Due to the ‘reactive’ (as opposed to proactive) nature of the data collection, it is not easy to create a general measure of primary care service utilisation using the GP data. Summing the dates patients have data recorded rather than the multiple records generated on each date, effectively sums up GP practice interactions. This is not ideal because it includes dates that were generated by administrative processes going on in the practice such as recording the information received from discharge letters etc. Further potential refinements to develop a robust means of estimating primary care activity are included in Appendix 5.

3.16 A small ‘snapshot’ of data for a fictitious but typical patient record is shown in the box below. It illustrates the potential complexity of the data recorded in primary care. Using the primary care event data often requires drawing inferences from patterns of ‘events’ that cannot be verified. For the purposes of the Project some regularly occurring codes have been utilised as indicators of patient activity. Some examples of these are described in the following paragraphs. The potential number of ‘patient pathway’ indicators is large and the selection of each one requires an examination of which codes are regularly used by GPs, how frequently, how completely the indicator gets recorded (for all patients or just some), and how the recording has changed over time. The examples included indicate some of the considerations that must be taken into account when choosing indicator variables from this dataset.

Tom Brown's knee: fictitious data based on real recording in SAIL

Tom Brown has five records in the GP event data between 4/09/2002 and 9/09/2002. The first two indicate that he had a "complete knee replacement using cement" and "manipulation of the knee", both on 4th September. The next three records indicate that on the 8th September he was prescribed pain killers, and two 'breathe easy 'aerosol inhalers. Then the final record on the 9th indicates he had an influenza vaccination.

This illustrates several features of the primary care database

- Many records can be generated on the same day, one record per 'event'.
- Records do not necessarily originate in the General Practice, as with the knee replacement, this would have arrived here via a discharge letter from the hospital that carried out the procedure.
- Manipulation of the knee may not be related to the same knee that was operated on. It may have been a check on Tom's other knee. In general it is not always possible to assume that information recorded on the same date refers to exactly the same condition.
- Records generated around the same time are not always related. Tom was being treated for the flu perhaps. Whether the painkillers for the recovering knee or the flu is unknown, and the Influenza Vaccination on the 9th September may have been scheduled prior to the prescribing events of the 8th September so are coincidentally recorded next to each other chronologically in the database.
- When the data are chronologically presented, the total history of Tom's knee is not neatly captured in one place – a further search revealed he first complained of knee pain in April 1999.

Using specific prescribing patterns as an indicator

3.17 The items prescribed during primary care are very precisely and consistently recorded by GPs, using codes in the Drug and Appliance Products section of the Read coding system[8]. There are thousands of drug codes and patient records show changes in drugs and dosages.

3.18 Numbers of prescriptions of drugs in three broad categories have been used in this Project - Lipid lowering drugs, anticoagulant drugs and Nicotine Replacement Therapy drugs. The more specific individual drug codes (for example 'bxd2.' identifies Simvastatin prescribed as 20mg tablets) have therefore been grouped. So, selection on the first two digits of the code ('bx' for Lipid lowering drugs), allows a broader category of drugs to be included.

Body Mass Index (BMI) and weight recording

3.19 Being overweight is known to be a risk factor for many diseases so it was expected that there would be some variation in health outcomes related to the weight or BMI of the patient. A recording of overweight or high BMI in primary care was therefore used as an indicator of overweight in the Project. However there are two sets of Read codes in use for capturing this type of information. Read codes starting '22K' refer to the recording of BMI with additional detail coded in the fourth and fifth digits of the codes, (e.g. '22K5.' Means BMI 30+). Read codes starting '22A' record weight taken on examination (e.g. '22A5.' Indicates weight greater than 20% over the ideal). Some GPs will use one set of codes and some another, and for some patients we have recordings of both codes. For the Project we combined this recording as shown in Table 3.2, below.

3.20 This recording has been used in two ways. The actual record of the weight and BMI is an important variable but the frequency with which measurements have been taken can also indicate the level of primary care involvement in addressing weight issues. For the purposes of this Report, we have calculated the recording frequency; the most up-to-date assessment of BMI was also captured, so this variable will be available for further analysis if required.

[8] For further information see http://systems.hscic.gov.uk/data/uktc/readcodes

Table 3.2 Grouped Read codes indicating weight or BMI recording

Read codes	Description
'22K1','22A2','22A6'	BMI low,10-20% below normal, underweight
'22K2','22A5'	BMI high, > 20% above ideal
'22K3','22A3'	BMI Normal, within 10% of ideal
'22K4','22A4','22AA'	BMI 25-29, 10-20% over ideal, overweight
'22K5','22A5'	BMI > 30, > 20% over ideal weight
'22K6','22A1'	BMI < 20,> 20% below ideal weight
'22K7'	BMI 40+
'22K8'	BMI 20-24

3.21 Sometimes the codes 22A and 22K appear in the data without the detail of the other two digits. In these cases sometimes, but not always, a numeric value is recorded in the same field as the code, which indicates an actual weight or BMI measurement. There may also be a height recording in the data, so this could be used alongside weight to calculate a BMI. A weight or BMI recording on their own may not indicate whether a person is overweight, since neither measurement distinguishes between fat mass and muscle mass. In a live GP system there is the facility to record a ‘free text’ entry, included alongside the chosen codes. The GP may include relevant notes next to each code, but free text data does not get transferred into SAIL due to its potential to include identifiable data. However, at the population level and therefore for this Project, it is safe to assume that the vast majority of weight and BMI measurements have been recorded in relation to overweight.

3.22 Other codes appear within the hierarchy that also imply a weight problem, such as ‘Health education – weight management’ and ‘weight loss diet’. Further work possible to refine the use of data on weight or BMI as an indicator is described in Appendix 5.

Smoking

3.23 Smoking is a serious risk factor in many diseases, so that there are likely to be different patient pathways for those who smoke and those who do not. Evidence that suggests the patient was being advised about smoking, or supported in giving up may therefore be informative for this Project. Capturing information about patients’ smoking status is quite complicated in the GP data.

Specific detailed codes need to be carefully grouped. For this report, we have grouped cases according to the categories 'non-smoker', 'smoker' and 'ex-smoker'. A list of 'smoking status' Read codes is included in Appendix 3. Some codes imply rather than confirm smoking status, such as 'not a passive smoker'. For the Project we have assumed that this is a non-smoker who is also not exposed to an environment where passive smoking regularly occurs.

3.24 Both the smoking status of the individual and the frequency of recording of smoking status over time were identified as possible indicators for use in the Project, the first allowing the effect of being a smoker to be taken into account, and the second as an indicator of how actively the GP practice was encouraging a behavioural change. We have explored the use of both of these indicators for the Project.

Hypertension

3.25 Blood pressure is an important diagnostic measurement in both Primary and Secondary care, as high blood pressure is often an early sign of developing problems, and can indicate anything from stress to circulatory disease. There are many Read codes that indicate hypertension, either in the individual or as a family history. Read codes starting with '246' are related to blood pressure recording, and the fourth and fifth digits of the code can indicate hypertension or hypotension or numerous other details, a list of which is included in Appendix 4. However, there are a great many records where just the '246' code is recorded. Some of these have a corresponding blood pressure reading recorded in the same field, but many do not. As for weight, there may have been other information recorded as free text and therefore not included in SAIL. It may be that '246' without further detail indicates that a blood pressure measurement occurred and was found to be in the normal range, but this can only be inferred.

3.26 As with smoking, the regularity of recording is indicative of the level of patient monitoring or management in Primary care for this condition. For the purposes of this Project, we calculated the number of BP measurement records as an indicator of the level of primary care monitoring, but due to the large number of non-specific BP codes, we also used specific Hypertension codes to indicate a patient diagnosed with hypertension.

Hospital in-patient data – the Patient Episode Database for Wales (PEDW)

3.27 PEDW is an all-Wales database containing records for in-patient or day case care carried out in Wales, plus treatments carried out on Welsh residents elsewhere in the UK. NHS Wales Local Health Boards are required to download, on a monthly basis, very clearly defined and standardised data from all hospital Patient Administration Systems (PAS). These are collated by the NHS Wales Informatics Service (NWIS), who also receive details of Welsh patients treated in England through a mechanism known as the NHS switching service.

3.28 Capturing data on the daily stream of patients entering and leaving hospitals throughout Wales begins with the collation of information from Hospital Patient Administration Systems. There have been very clearly defined data recording standards in place since around 1999, and all hospital activity on a day case or inpatient basis is regularly submitted into NWIS for inclusion in PEDW. The dataset is anonymised into SAIL which currently holds details of more than 2.8 million patients receiving approaching 18 million spells of care. There is a great deal of coded detail about each patient interaction, with only 15% of patients having only a single diagnostic code recorded during a spell of care, and almost half of all patients having 5 or more diagnoses recorded. The vast number of diagnosis codes means that there are thousands of ways to group patients with different sets of co-morbidities.

3.29 PEDW contains ‘finished consultant episodes’ of care. A ‘finished consultant episode’ is defined as a completed ‘unit’ of care under the care of one consultant. Each episode has provision for a number of diagnosis and operative procedure codes to be recorded. In PEDW, the ICD 10 diagnostic codes are used. So, for example, first episodes of care containing a diagnosis in the range I00-I99 relate to episodes of cardiovascular diseases. A typical set of PEDW records is described in the box below.

Mrs Smart's Stroke: fictitious but based on realistic PEDW data

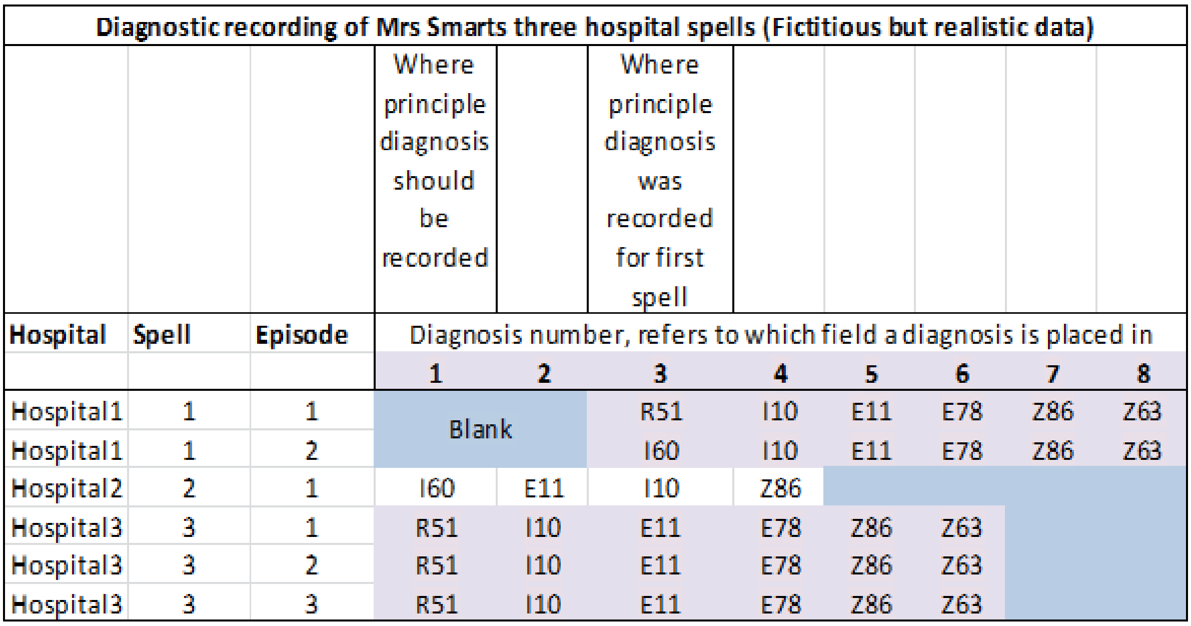
Mrs Smart was admitted on 22/10/2009 as an emergency at the request of her GP. She was 79 and her birthday was the following day. Her admission was coded under Specialty 'Geriatric Medicine', and she remained in hospital until the 27/10/2009. She was first treated by consultant A, a 'General Medicine' specialist. Six different diagnostic codes were later coded from Mr A's notes , indicating that she was admitted with a Cerebral Infarction (stroke) but also suffering from Hypertension, Non-insulin dependent diabetes, a history of diseases of the nervous system and a history of diseases of the circulatory system, and that she had a dependent relative needing care at home. This took less than 1 day and the episode length was recorded as 0 days. One operative procedure was recorded at this time (a U05 which indicates that some kind of examination took place). Then she was looked after in high dependency bed managed by the team of consultant B, a geriatric specialist for 1 day, and then she was moved to more general ward for the next four days looked after by consultant C who is also a geriatric specialist. She then left hospital on 27/10/2009 aged 80. She was discharged to her own home.

Each time she was under the care of a different consultant is classed as what is called a finished consultant episode. Mrs Smart had 3 episodes. The time she was continually in hospital is known as a 'spell' of care. This describes one 3 episode spell of care lasting 5 days.

All this information was either recorded as a code or a date, or could be calculated from the other data. E.g. age can be calculated based on the date of birth and the admission or discharge date. So if calculated on admission date she was 79, but if calculated on discharge date she would be 80, and if being grouped into five year age bands this choice would place her in a different age group.

3.30 Each row of data recorded in PEDW represents an episode of care, and there are 14 columns to allow different diagnoses to be coded. Column 1 of the diagnosis columns is where the diagnosis that led to the admission is normally placed. This is usually referred to as the 'primary diagnosis', but as we also discuss primary care records in this Project we will use the term 'principal diagnosis'. Examination of Mrs Smart's PEDW records revealed that she had been admitted twice previously, and to two different hospitals. The diagnosis recordings from one hospital do not start in column 1 of the diagnosis fields. This is illustrated in Figure 3.2, below. The principal diagnosis recorded for this spell was in the third diagnosis column. This is a common problem affecting 2% of records, and believed to be caused by the outputs from some older PAS systems. As a consequence, when searching records in the database by principal diagnosis, it is important not to search only the first column.

Figure 3.2 Misplaced recording of the principle diagnosis



Diagnostic recording of Mrs Smarts three hospital spells (Fictitious but realistic data)										
			Where principle diagnosis should be recorded		Where principle diagnosis was recorded for first spell					
Hospital	Spell	Episode	Diagnosis number, refers to which field a diagnosis is placed in							
			1	2	3	4	5	6	7	8
Hospital1	1	1	Blank		R51	I10	E11	E78	Z86	Z63
Hospital1	1	2			I60	I10	E11	E78	Z86	Z63
Hospital2	2	1	I60	E11	I10	Z86				
Hospital3	3	1	R51	I10	E11	E78	Z86	Z63		
Hospital3	3	2	R51	I10	E11	E78	Z86	Z63		
Hospital3	3	3	R51	I10	E11	E78	Z86	Z63		

Office for National Statistics (ONS) Mortality Data

3.31 Death data from ONS has been anonymised into SAIL for the years 2003 to 2011. The data include date of death and cause of death. Through data linking, it is possible to establish the cause of death that was recorded for those patients included in the Project who had died before 31.12.2011. For the purposes of this Project, we are most interested in deaths from stroke but when examining other chronic conditions or combinations of conditions, stroke would be relevant either as a main condition or as a co-morbidity.

Deprivation

3.32 The strong association between health and deprivation is well known and documented. Deprivation may therefore be an important factor for consideration when establishing care pathways. The Welsh Index of Multiple

Deprivation (WIMD) ranks Welsh LSOAs according to area based deprivation indicators from a number of domains e.g. income and housing. The latest available version of the WIMD was released in 2011 and this has been used in the Project. The problem with using area based scores assigned at an individual level is that an average area score is assigned to the individuals when in reality they may be either relatively more or less deprived than the average for their neighbourhood. Using data linkage, a WIMD score was assigned to each member of the set of stroke patients based on the LSOA where they were living when the stroke occurred. Further refinements to the use of this data are described in Appendix 5.

4 Methodology

4.1 The original specification for this project included the development of patient pathways for patients with a number of chronic conditions. The initial exploration of the data indicated that the number of variables involved and consequently the vast numbers of potential patient groups that could be examined was far more than could be handled within the scope of a demonstration project. It was therefore agreed with WG policy leads to concentrate the analysis on individuals who had experienced a stroke and individuals with diabetes and to develop a methodology that could potentially be used for other cause groups in the future.

4.2 The methodology described below was developed for stroke patients and, within that set, includes patients with diabetes identified as co-morbidity.

4.3 The key methodological challenges of this project were to:

- Develop a way of selecting a set of ‘similar patients’ from very large numbers of patients with multiple chronic conditions.
- Find a way to align the set of patients according to the stage of the condition so that their pathways would be comparable.
- Determine ‘care pathway’ milestone indicators on which to compare them.
- Determine how to group patients according to the similarities identified using the indicators.
- Explore how patient treatment pathways differ, and why.

Initial selection of patients

4.4 The raw data related to patients who were at different stages of disease at different times so, in order to compare them, it was necessary to align the individual timelines of stroke events in some way. A two-stage process was adopted. Firstly, in order to establish a set of stroke patients (the Project Set), all patients with an emergency admission with the principle diagnosis ‘Stroke’, between the dates of 01.01.2007 and 31.12.2009 were selected. For most of the Project Set, this would be the first emergency stroke admission but for others it would be a second or subsequent stoke. The second step involved looking back through the PEDW data to establish the date of each patient’s

First Emergency Admission for Stroke (FEAS). The FEAS date was used to align all the patient timelines in order to compare health service events when patients would be at a similar stage in their disease. Whenever their first stroke happened in history (from when PEDW data is available from 01.01.1999 to 31.12.2009), 'care pathway' events for the set of patients were compared in relation to the timing of their first stroke.

Development of potential care pathway indicators

4.5 The Project sought to establish which records in a patient's history might make suitable patient pathway indicators. A wide range of variables were considered. To compare the care history of each patient, both secondary and primary care activity was captured where possible. Secondary care activity before the first admission for a stroke was identified for the full set of stroke patients by looking back at the previous admissions for any cause - both emergency and elective - as recorded in the PEDW database, for the four years prior to the FEAS.

4.6 As noted above, full primary care data is only available in SAIL for around 47% of residents of Wales. Those stroke patients for whom data on primary care activity was available were flagged through data linkage to make it simple to identify them for analysis purposes.

4.7 Good quality primary and secondary care data are available from about the year 2000 onwards in SAIL. By inspection of the FEAS dates it was established that data of a consistent quality would be available for all the selected patients for up to four years prior to the FEAS. Records for each patient from primary and secondary care datasets were included in the Project using their individually assigned start dates, which were therefore calculated as (4x365=) 1460 days before their FEAS.

4.8 The Project investigated the care history of each patient. This is only complete for patients who were permanently living and registered in Wales during the Project time period. To establish what proportion of the care history was available for patients who migrated in or out of Wales during the four years prior to their FEAS, we counted the number of days each patient was registered with a SAIL-participating GP practice in the four years before the FEAS date. We used this information to create a variable in the dataset that could be used to calculate ‘exposure periods’ in ‘person days’ in any further work; this would, for example, allow the inclusion of individuals with a patient history of less than 2 years available. Appendix 5 includes suggested uses of this alternative methodology.

4.9 In order to explore what information recorded in the care history might prove useful in defining care pathways, a number of data items and derived indicator were calculated using the primary and secondary care data (Table 4.1). Information was included for the 4-year block of time prior to the FEAS.

4.10 The items asterisked in Tables 4.1 to 4.3, below, are those used in the cluster analysis.

Table 4.1 List of health data and indicators used in the analysis: before stroke

No.	Indicator description	Units
1*	The number of emergency admissions (for any cause)	Number
2	The number of elective admissions (any cause)	Number
3	Sum of days spent in hospital for all previous admissions (any cause)	Days
4*	Sum of days spent in hospital for all previous emergency admissions (any cause)	Days
5*	A list of co-morbidities recorded in Secondary care (e.g. used to include diabetics in the cluster analysis)	Coded list
6*	Total number of GP event dates (where at least one record was recorded at the GP practice)	Number
7	Sum of time registered with a participating GP practice	Days
8*	Number of blood test result events	Number
9*	Number of blood pressure recording events	Number
10*	Number of smoking status recordings (indicative of extent to which smoking habit was discussed with primary care practitioners)	Number
11*	Number of statin prescriptions	Number
12*	Number of anti-coagulant prescriptions	Number
13*	Presence of weight or BMI measurement activity record	Yes or No
14	Weight or BMI measurement recorded closest to FEAS in date	Coded
15	Smoking status recorded closest to the FEAS in date	Coded
16	Prescription of nicotine replacement therapy	Yes or No
17	Patient classified with high blood pressure (hypertension)	Yes or No

* Indicators used in the cluster analysis.

4.11 A set of socio-demographic variables were also captured or calculated for each patient on the FEAS date (see Table 4.3, below)

Table 4.2 List of socio-demographic characteristics used in the analysis

No.	Indicator	Units
18	Age Group	Years
19*	WIMD tenth/decile, based on residence (coded 1 to 10†)	WIMD
20*	Area type (Urban=1, Town & Fringe=2, Village, Hamlet & Isolated Dwellings=3)[9]	Code
21	Gender code (1=Male,2=Female)	Code

* Indicators used in the cluster analysis. († where 1 = most deprived)

Post-stroke health service activity

4.12 Comparing activity following the FEAS date has to take into account that those who had a FEAS date at the end of 2009 have only had the opportunity to survive two years to the end of 2011, whereas those having earlier FEAS dates have had a the opportunity to survive for longer. Taking 31.12.2011 as a cut-off date, the maximum time span for which we can compare activity is two years.

4.13 Table 4.3 below describes variables that were identified at individual patient level for the two year period after the FEAS date.

[9] The standard ONS Rural/Urban Definition 2004 was used: Urban (population over 10,000), Town and Fringe and combined 'Village' and 'Hamlet and Isolated Dwellings' (see http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/rural-urban-definition-and-la/rural-urban-definition--england-and-wales-/index.html)

Table 4.3 List of data and health indicators used in the analysis: after stroke

No.	Indicator	Units
22*	Survival (to death or to 2 years following FEAS)	Days
23	The number of emergency admissions (for any cause)	Number
24	The number of elective admissions (any cause)	Number
25	Sum total of all days spent in hospital	Days
26	Sum total of all days spent in hospital for emergency admissions	Days
27	A list of co-morbidities recorded in Secondary care	Coded list
28	Total number of GP event dates (where at least one record was recorded at the GP practice)	Number
29	Sum of time registered with a participating GP practice	Days
30	Number of blood test result events	Number
31	Number of blood pressure recording events	Number
32	Number of smoking status recordings (indicative of how much smoking habit was discussed with GP practice)	Number
33	Number of statin prescriptions	Number
34	Number of anti-coagulant prescriptions	Number
35	Presence of weight or BMI measurement activity recorded	Yes or No
36	Weight or BMI measurement closest to Final* date	Coded
37	Smoking status closest to the Final* date	Coded
38	Prescription of nicotine replacement therapy	Yes or No
39	Patient classified with high blood pressure (hypertension)	Yes or No

* Indicators used in the cluster analysis.

The characteristics of the stroke patient population

4.14 Descriptive statistics were produced for the set of stroke patients. These were inspected to establish if obvious subgroups of similar patients could be identified. A large number of potential subgroups were identified based on age group, gender, type of stroke, and socio-economic group. Within each subgroup, very large numbers of co-morbidities are recorded in secondary care. There are potentially 1000s of separate care pathways that could be derived from the data.

4.15 To try to reduce the dimensions of the data, a statistical technique called ‘clustering’ or ‘segmentation’ was used try to identify some specific stroke patient ‘types’. The presence of diabetes, by diabetes type, was included in the analysis as an example of a co-morbidity that might generate different treatment or care pathway events.

Analysis of Mortality

4.16 ONS mortality data was linked to the records for the set of stroke patients to establish cause of death for those who had died by 31.12.2011. This data was used to confirm the date of death, and to measure the extent to which the selection method, (based on secondary care diagnosis), was missing stroke patients. A frequency table of deaths by cause was produced for the non-surviving members of the Project Set in order to establish the extent to which strokes were recorded as the main cause of death for these people.

The effect of diabetes as a pre-existing co-morbidity

4.17 As noted above, the initial intention was to analyse health records for patients with multiple chronic conditions. Having reduced the scope of the project, we chose to focus on stroke patients with diabetes as a co-morbidity. For the purposes of this Project, we therefore added a flag to the set of stroke patients that identified those who were diabetic and used Item 4 from Table 4.2 (the list of co-morbidities) to identify the type of diabetes with which each of these patients had been diagnosed. Comparative statistics were produced for those people with diabetes and those with no diabetes to establish variations in hospital visits, length of stay and survival differences (see Chapter 7).

4.18 In future, the SQL code developed in order to complete this analysis for diabetes can be adapted relatively simply to complete this kind of comparison for other single co-morbidities and for multiple co-morbidities (see Appendix 5 for details of potential future analysis).

4.19 A key tool developed as a result of this Project is the creation of a single searchable data field for each patient that lists all co-morbidities recorded over time. This will facilitate all future analysis of linked health data in SAIL.

Exploration of a specific subgroup – young male stroke patients

4.20 Looking for patterns (or types of patient pathway) by doing detailed crosstabulation analysis of the whole Project Set of stroke patients would have been impractical because of the size and heterogeneity of the dataset i.e. because we would have been unable to make sense of such a vast amount of variation using such simple, bivariate analysis methods. However, some exploratory cross-tabulation analysis was necessary to further investigate the suitability of the chosen indicators. Within the limited scope of the demonstration project, it was therefore decided to focus on a subgroup of patients who might have relatively little co-morbidity but for whom there were nevertheless sufficient numbers of patients to allow robust analysis of the chosen indicators. ‘Younger’ male stroke patients (aged under 40 years) were chosen since this group had relatively few co-morbidities (much co-morbidity is age related) but (unlike ‘younger’ women) there were sufficient numbers to attempt to establish some similarities in patient pathway.

Cluster Analysis

4.21 A statistical technique called cluster analysis was used experimentally to split the stroke patients into groups of similar individuals, based on the asterisked indicators listed in Tables 4.1 to 4.3, above. The indicators were chosen on the basis that they were most likely to be related to stroke. Future analyses could include further indicators of health service use. Where the initial intention was to examine the complexity of health service use for patients with multiple chronic health conditions, additional co-morbidities could be included in future studies as well as additional indicators relating to those co-morbidities e.g. prescribing activity for additional drugs.

4.22 There are two main methods of cluster analysis: agglomerative hierarchical clustering and k-means clustering. Each method has advantages and disadvantages.

4.23 The agglomerative hierarchical approach has the main advantage that it can be used to identify the ideal number of meaningful clusters that can be created based on the available data; however, the major drawback is that once an individual is assigned to a cluster, they cannot be re-assigned even if the cluster takes on very different characteristics as it grows.

4.24 The main advantage of the k-means method is that is it very simple and quick both to use and to interpret; however, two concerns with the method are that a) the number of clusters must be chosen in advance, and b) the final cluster solution can be dependent on the original starting cluster centres used.

4.25 Given the limited scope of the demonstration project, we chose to use the k-means method to identify cluster solutions that minimised the within-cluster differences between patients and maximised the between cluster differences[10]. Using the SPSS statistical analysis package, we requested solutions generating between 7 and 10 clusters in order to evaluate each solution. The most appropriate solution was selected by examining the extent to which the memberships of each cluster were homogenous in terms of the characteristics of the stroke patients.

4.26 Within the limited scope of a demonstration project, the clustering method is being used experimentally and purely to illustrate the kinds of analysis that might be used to identify patient pathways. In future, when more detailed discussion with policy and analytical colleagues has identified both the health indicators and the combinations of conditions that are of most interest, more detailed statistical methods could be used to further explore the robustness of the selected solutions. For now, the findings, which are presented in Chapter 8, are purely for illustrative purposes.

4.27 Future projects might consider using a combination of the two clustering methods, using the hierarchical approach to identify the optimal number of clusters then the k-means approach to identify the most robust solution selecting that number of clusters[11].

4.28 Once the clusters had been identified, the clusters or groups of patients were profiled in terms of both the health service indicators and the socio-

[10] This was done using the Euclidean Sum of Squares (ESS).

[11] To establish the optimal solution, the variance in the indicators explained by the solution (a) the F-statistic and b) the Between Cluster Sum of Squares) could be plotted against the number of clusters. This is known as the 'Elbow' method.

demographic variables that were excluded from the cluster analysis e.g. age, gender, urban-rural location and the Welsh Index of Multiple Deprivation. The aim of this process was to examine whether meaningful clusters had been generated both in terms of patients' health service use before the stroke and their socio-demographic characteristics.

4.29 For the purposes of the cluster analysis, where we were seeking to identify groups of patients with similar pre-stroke health service histories, it was important to use a range of both primary and secondary care information. Stroke patients were therefore only included when at least two complete years of health data was available prior to the FEAS. As noted above, primary care event data is only available for around 47% of patients, so this reduced the numbers of stroke patients included in the analysis from 18,744 to 8,781. Of these, 6,928 (37% of the Project Set) had at least two complete years of health data available.

4.30 In preparation for the clustering, variables with large numbers of values (e.g. 'Total days in hospital prior to the FEAS', which ranged from 0 to 1460) were summarised into variables with between 5 and 8 categories. Because the clustering process uses the amount of variation within a variable to assign cluster membership, using categorical variables with relatively few categories will give them a disproportionate influence on the analysis; it is therefore better to create variables with relatively similar numbers of categories. The way variables were categorised was based on a visual inspection of the frequency with which different values occurred, creating categories that reflected the main variations seen in the data. One variable, indicating whether the GP had measured weight or BMI, had only two categories, indicating 'yes' or 'no'; it was unfortunate that the weight / BMI measurement could not be used but in its absence, it was decided the variable represented a sufficiently important association with stroke to include it, despite the fact that a binary variable would have slightly disproportionate weight in the cluster analysis. The different types of stroke were grouped into haemorrhage, infarction, 'undetermined whether haemorrhage or infarction', transient cerebral ischaemic attack and 'other stroke'; retinal vascular occlusions were excluded due to the very small number recorded and the very different nature of this condition in comparison with other strokes (see Table 5.1, below, for the full list of stroke types). Further details of the variables created for the cluster analysis are included in Appendix 6.

4.31 A set of health service use indicators was included in a k-means cluster analysis in the SPSS Statistical Package to establish a number of distinct clusters or groups of stroke patients. As noted above, diabetes was included as a co-morbidity. For the purposes of this experimental use of cluster analysis, type of stroke was included as a variable to examine whether this contributed to the understanding of the clusters; future cluster analysis could be done for each type of stroke separately.

4.32 A separate technical guide will be made available via SAIL to include the SQL coding developed for the project and detail of the variables utilised to develop indicators for the Project.

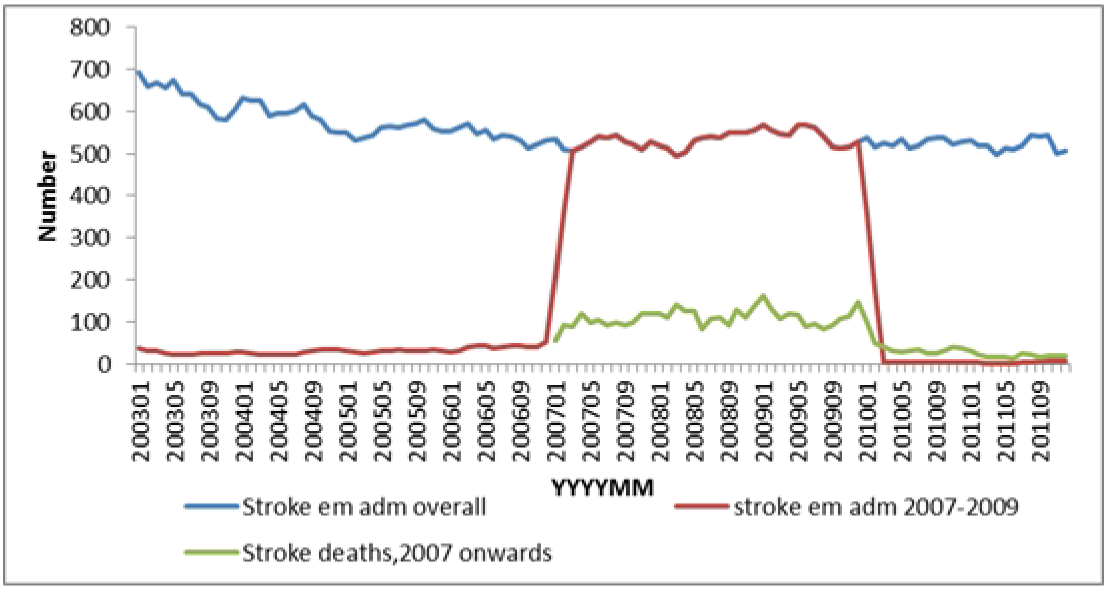
5 The Characteristics of Patients having emergency admissions for Stroke 2007-09

5.1 As noted above in Chapter 1, the reader should bear in mind that the Project was designed to demonstrate the usefulness of linked data in allowing patient pathways to be developed; findings that relate to health service use are therefore presented as steps along the way to achieving that objective, not to serve as a report on the health of the population of Wales.

5.2 The selection process described in Chapter 4, above, identified a set of 18,744 individuals who were admitted between 01.01.2007 and the 31.12.2009 and where the principal diagnosis was a stroke. The graph below (Figure 5.1) shows the number of emergency admissions per month for the general population. The Blue line in Figure 5.1 indicates the emergency stroke admissions for the whole population across the whole Project period of 2003 to 2011. The red line indicates the emergency admissions that relate only to the Project Set of patients, i.e. the sub-set of patients who had stroke admissions between 2007 and 2009; this is why the vast majority of strokes for the Project Set occur in those years. However, these admissions are not necessarily for the first strokes these patients have had. So, looking only at the Project Set, we also looked back through their patient histories in order to identify the *first* stroke experienced by those patients. The red line therefore shows that small numbers of strokes had happened before the selection period began on 01.01.2007. This indicates that the majority of the strokes occurring to our Project Set of patients were first strokes but a small number were second or subsequent strokes. This is why the ‘FEAS’ (first emergency admission for stroke) date was used as a benchmark event to ensure that we were comparing patient histories at a comparable stage i.e. their first stroke.

5.3 It is also clear from the red line in Figure 5.1 that the Project Set continued to have stroke-related emergency admissions after the end of 2009; so, these are all re-admissions for subsequent strokes occurring after the selection period ends at 31.12.2009.

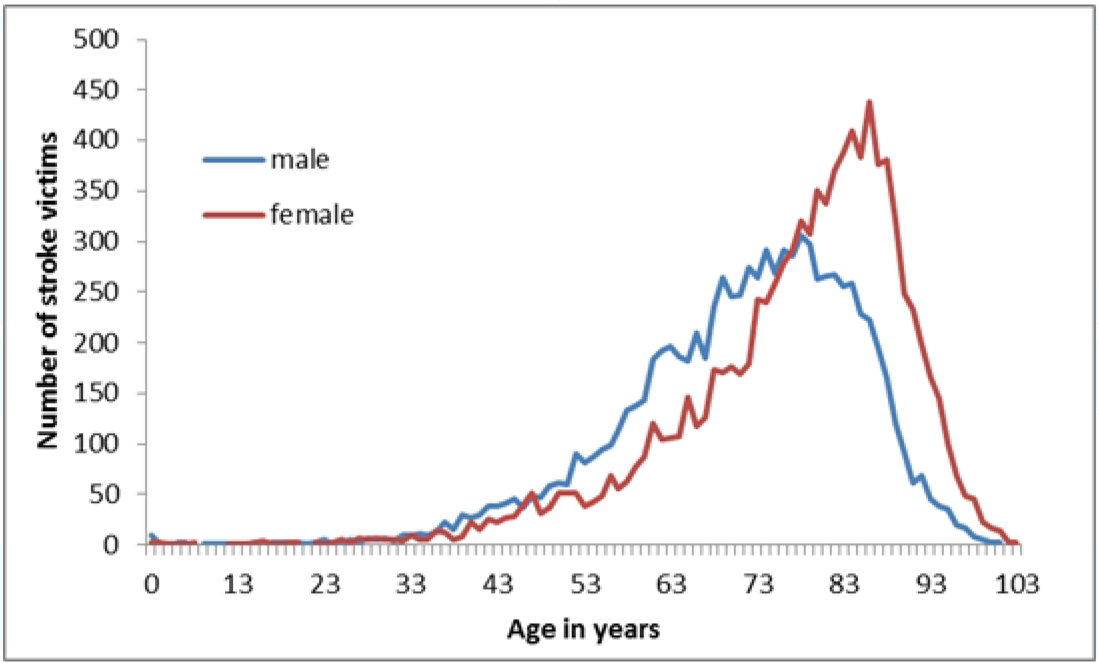
Figure 5.1 Number of emergency stroke admissions in Wales from 01.01.2003 to 31.12.2011, overlaid with a) admissions to project set and b) stroke deaths to project set from 2007



5.4 The Project Set was made up of 8,962 male and 9,782 female patients across the complete age range from birth onwards. Figure 5.2, below, shows the distribution of these cases by age and gender. Significantly more men than women had strokes before the age of 70 years. This finding is in line with previously published information about the characteristics of stroke patients in Wales[12]. Among those aged of 70 years and above, a significantly greater number of women than men had strokes. For men, the maximum frequency of emergency stroke events occurred at the age of 78 years whereas for females this occurred at 86 years. From these distributions, it appears that the majority of strokes occurred over the ages of 65 years and 72 years in men and women respectively.

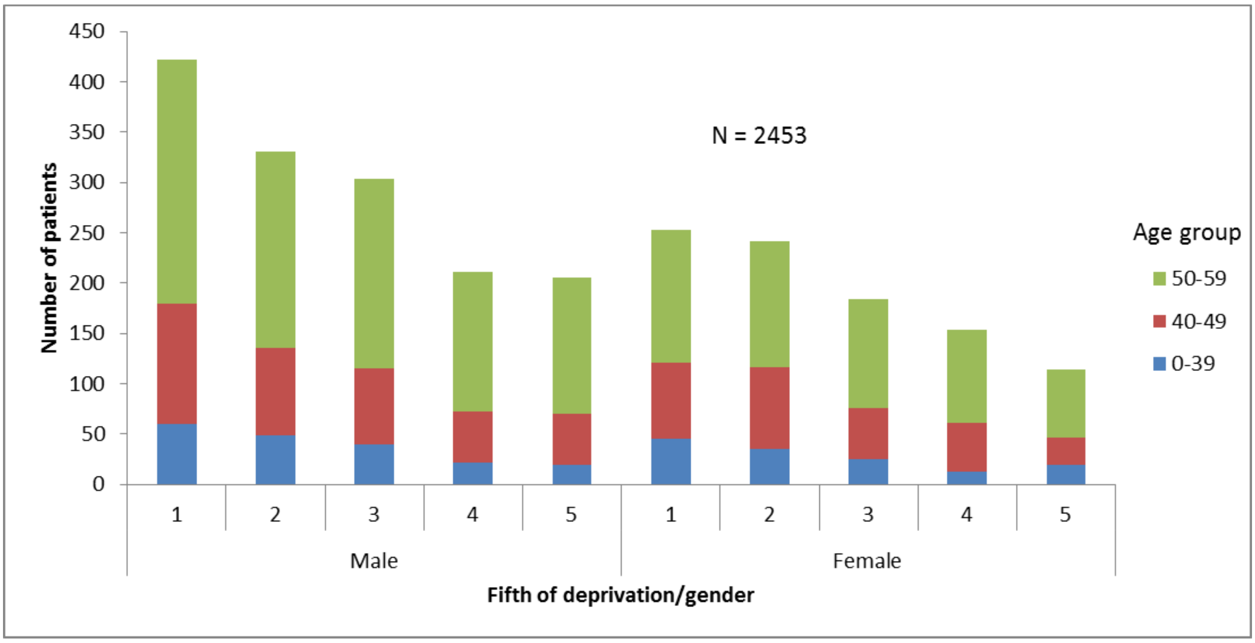
[12] http://www.wales.nhs.uk/sitesplus/888/page/44358

Figure 5.2: Age and gender distribution of Stroke Emergency admissions.



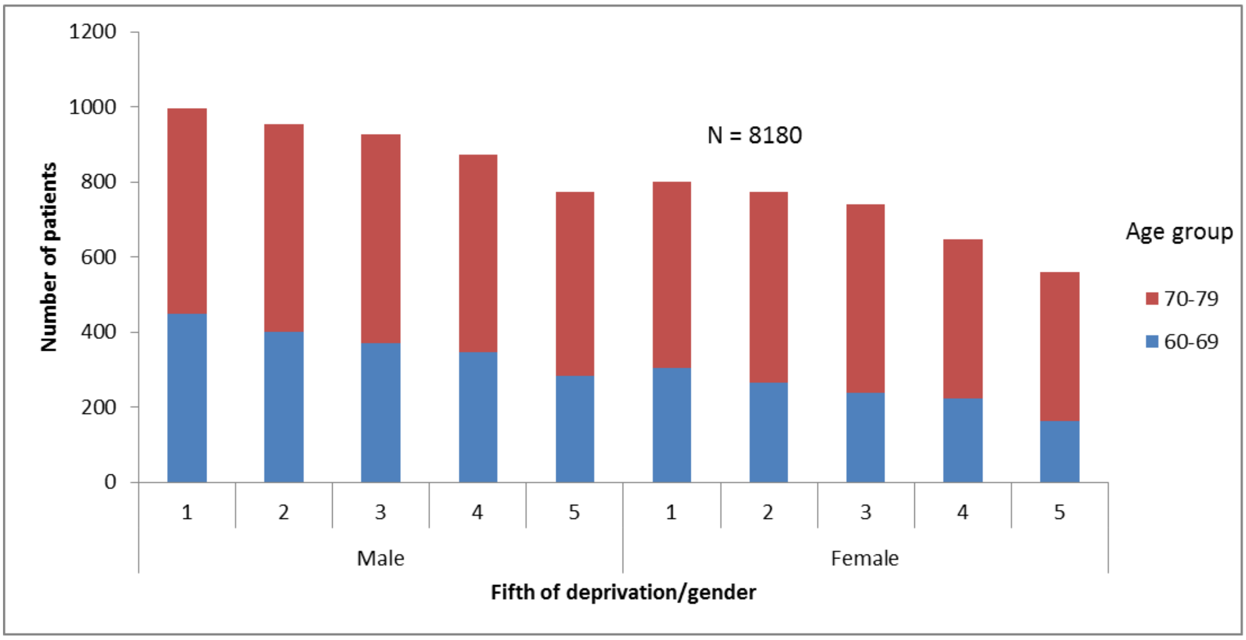
5.5 Figures 5.3 a, b and c below show the number of emergency admissions divided by age group, gender and deprivation fifth (as noted above, using fifths of WIMD). Figure 5.3a shows the numbers of admissions occurring in all people aged less than 60 years; there were greater numbers of men than women. For both genders, numbers of stroke admissions increased as deprivation increased.

Figure 5.3: Stroke emergency admissions by age group, gender and deprivation for people aged 0-59 years (1 indicates most deprived)



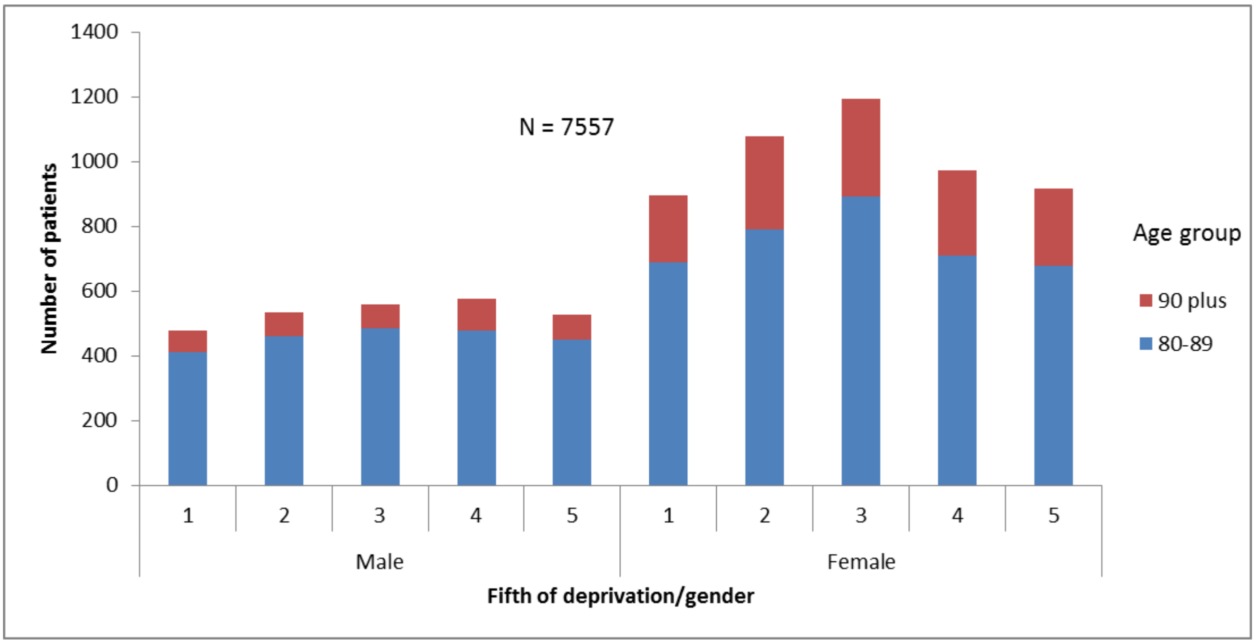
5.6 Figure 5.3b, below, shows admissions for people aged 60-79 years; there were greater numbers of men than women. For both genders, significantly greater numbers of stroke admissions occurred in the more deprived groups of patients. .

Figure 5.3b: Stroke emergency admissions by age group, gender and deprivation for people aged 60-79 years



5.7 Figure 5.3c, below, shows admissions for people aged 80 years and over; there were greater numbers of women than men. This reflects the greater life expectancy of women - more women than men survive to this age. For men, there were more admissions in the less deprived groups than in the more deprived groups, which also reflects survival - fewer men survived to this age in the more deprived groups. For elderly women, the numbers being admitted were less clearly defined by deprivation category.

Figure 5.3c: Stroke emergency admissions by age group, gender and deprivation for people aged 80 years and over



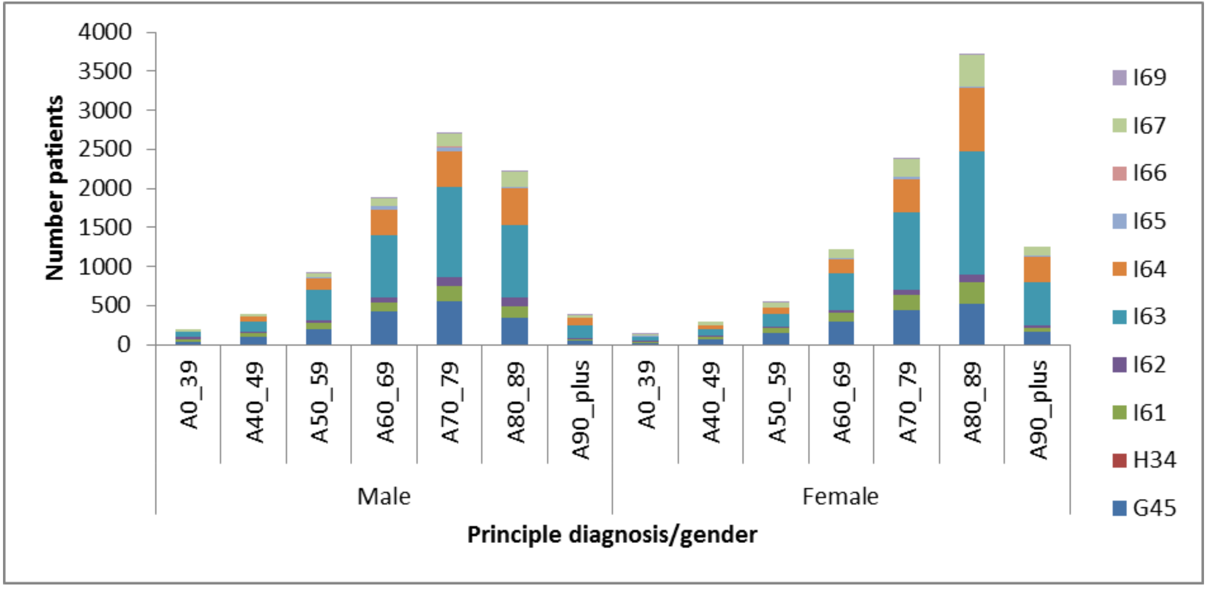
5.8 Strokes fall into the category of diseases known as ‘cerebrovascular disease’ (ICD 10 I61 to I69). Descriptions of the type of disease that fall into this category are shown in Table 5.1, below.

Table 5.1 ICD 10 codes and descriptions for Stroke types

ICD 10 Code	Description
I161	Intra-cerebral haemorrhage
I162	Other non-traumatic intracranial haemorrhage
I163	Cerebral infarction
I164	Stroke, not specified as haemorrhage or infarction
I165	Occlusion and stenosis of pre-cerebral arteries, not resulting in cerebral infarction
I166	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
I167	Other cerebrovascular diseases
I168	Cerebrovascular disorders in diseases classified elsewhere
I169	Sequelae of cerebrovascular disease
H34	Retinal vascular occlusions
G45	Transient cerebral ischaemic attacks and related syndromes

5.9 Cerebrovascular emergency admissions by type, age group and gender are shown in Figure 5.4 below. The stroke type is as recorded at FEAS. The most frequently occurring cerebrovascular disease in the Project Set was ‘Cerebral Infarction’, followed by ‘Stroke, not specified as haemorrhage or infarction’ then by haemorrhagic type strokes. Whilst each coloured segment in Figure 5.4 represents a separate stroke type, possibly requiring a different patient pathway, the individual types were aggregated into 5 groups for the cluster analysis, as described in 4.31 above.

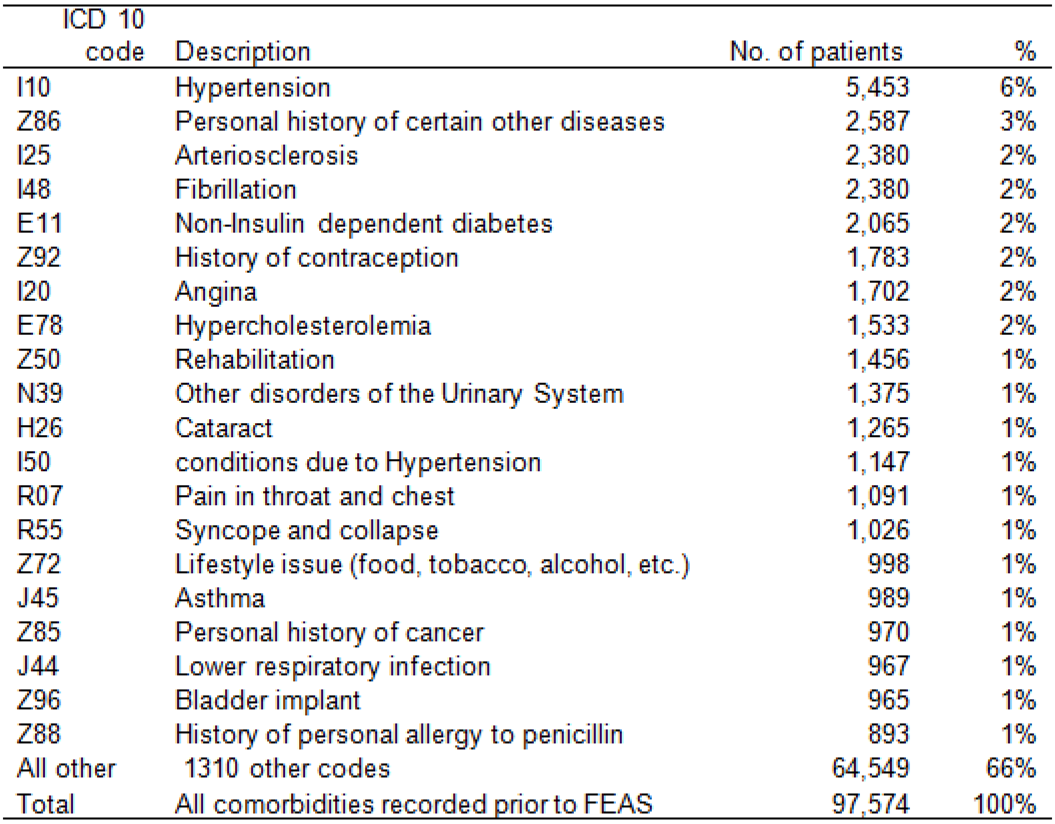
Figure 5.4: Stroke emergency admissions by age group, gender and type of stroke (N.B. the Stroke codes are explained in Table 5.1, above)



Co-morbidities occurring in stroke patients prior to the FEAS

5.10 During the four years prior to each person's FEAS, the overall average length of stay in hospital prior to and during FEAS was significantly longer for women (61 days, 95% C.I. 56.2 to 65.8) than for men (44 days, 95% C.I. 39.6 to 48.8). This may be explained by gender-related differences in age on admission and the associated co-morbidities. During these hospital visits, and in addition to stroke, numerous other diagnosis codes were recorded in secondary care (for a summary, see Table 5.2, below). A total of 1,330 diagnostic codes were used 97,574 times to describe conditions diagnosed for people in the Project Set of stroke patients. Table 5.2 lists the top twenty co-morbidities but these account for only 33% of the co-morbidities recorded. This makes the number of possible permutations of multiple co-morbidities very high, and therefore our likely ability to distinguish common patient pathways correspondingly low.

Table 5.2 Co-morbidities recorded in primary care during the four years prior to FEAS[13]



ICD 10 code	Description	No. of patients	%
I10	Hypertension	5,453	6%
Z86	Personal history of certain other diseases	2,587	3%
I25	Arteriosclerosis	2,380	2%
I48	Fibrillation	2,380	2%
E11	Non-Insulin dependent diabetes	2,065	2%
Z92	History of contraception	1,783	2%
I20	Angina	1,702	2%
E78	Hypercholesterolemia	1,533	2%
Z50	Rehabilitation	1,456	1%
N39	Other disorders of the Urinary System	1,375	1%
H26	Cataract	1,265	1%
I50	conditions due to Hypertension	1,147	1%
R07	Pain in throat and chest	1,091	1%
R55	Syncope and collapse	1,026	1%
Z72	Lifestyle issue (food, tobacco, alcohol, etc.)	998	1%
J45	Asthma	989	1%
Z85	Personal history of cancer	970	1%
J44	Lower respiratory infection	967	1%
Z96	Bladder implant	965	1%
Z88	History of personal allergy to penicillin	893	1%
All other	1310 other codes	64,549	66%
Total	All comorbidities recorded prior to FEAS	97,574	100%

5.11 The most commonly occurring co-morbidities were for what is classified in the ICD 10 coding scheme as ‘Factors influencing health status and contact with the health services’. Examples of these are items starting with a ‘Z’ in Table 5.2, above. Codes starting with ‘Z’ include the recording of historic health problems, follow-up and convalescence, and exposure to infectious diseases[5]. The next largest group of co-morbidities recorded for the Set of stroke patients related to ‘Circulatory Diseases’, such as hypertension and fibrillation. Many of these co-morbidities would affect how a patient is treated by health services, e.g. a history of adverse reactions to a particular type of drug or the presence of pre-existing hypertension.

5.12 From even this basic analysis, it can be seen that even a single individual’s health records are complex to analyse. Looking at numerous patients with a

[13] Z86 is the sum of a personal history of Z860 Neoplasms; Z861 Infectious and parasitic diseases; Z862 Diseases of Blood and blood forming organs; Z863 Endocrine nutritional and metabolic diseases; Z864 Psychoactive substance misuse; Z865 Mental and behavioural problems; Z866 Diseases of the nervous system; Z867 Circulatory diseases.

whole variety of potential co-morbidities creates a vast, almost endless matrix of potential condition combinations. Even excluding conditions/issues that would be theorised to have no bearing on strokes, a huge number of possible combinations remains. A key lesson learned as a result of the Project was that constraining the analysis was important in order to deliver something meaningful within a limited time scale. As noted above, although the initial intention was to focus on a range of conditions, we were forced to limit the focus of the project to stroke plus a single co-morbidity. After other circulatory disease conditions and 'history-of' diagnoses, the next most frequently occurring co-morbidity in stroke patients (Figure 5.2) was non insulin dependent diabetes. For this reason and due to interest from policy colleagues, diabetes was chosen as the condition with which to explore the methodological development. In future, and in discussion with the relevant policy and analytical colleagues in WG, analysis of further main conditions and co-morbidities could be completed. Further possible work is listed in Appendix 5.

Analysis of Mortality

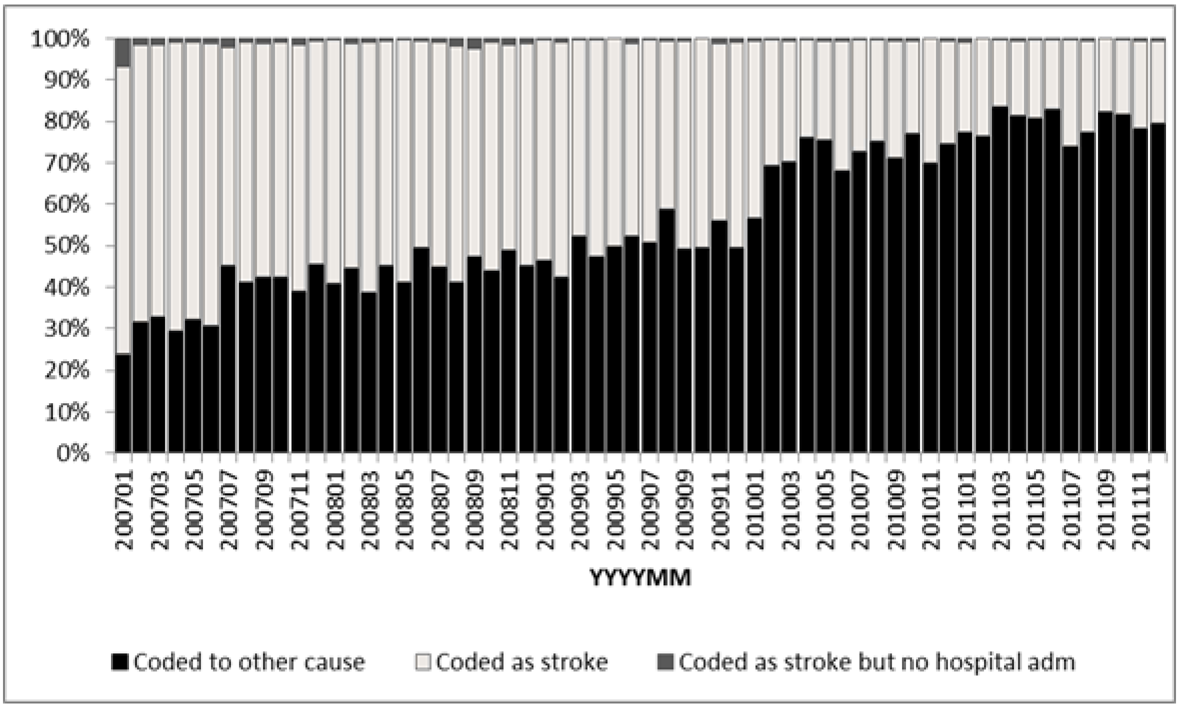
5.13 Many but not all of the strokes deaths occurring in the Project time period are recorded in the WDS database, and the WDS data set does not include a cause of death. However, ONS mortality data is routinely anonymised into SAIL, (i.e. is ALFed by NWIS through the same process as described for PEDW). Linking the set of stroke patients to ONS mortality data enabled the confirmation of date of death and allowed the identification of cause of death from death certification. The ONS utilise the same ICD 10 coding scheme as is used in PEDW.

5.14 The ONS mortality data included some deaths that are attributed to stroke but for which there is no match in the PEDW database. These cases are shown in green in Figure 5.5, below, and are individuals who had no interaction with secondary care during their stroke emergency but for whom the stroke proved fatal. There were on average three such cases per month in the period 01.01.2007 to 31.12.2009, a total of 120 fatalities. These cases have not been included in the analyses that follow but exploration of further linkage to determine whether these individuals have other primary and secondary care records is an area of potential further analysis included in Appendix 5.

5.15 Figure 5.5, below, shows the numbers of deaths by month for the set of stroke patients as recorded by ONS. The data are coloured according to whether the cause of death was a stroke (shown separately when identified using mortality data or from GP event data) or ‘any other cause’. In January 2007, 76% of deaths among the set of stroke patients were attributed to stroke. The proportion of deaths attributed to stroke then diminishes on a month by month basis. This is because the shorter the survival period following the stroke, the more likely it is that stroke will be recorded as the cause of death. The figures for February 2007 include the short-term stroke survivors for that month and deaths from the survivors from January 2007, and so on.

5.16 After December 2009, no new stroke patients are joining the set; so, while more time is passing between the FEAS and the date of death, the proportion of deaths attributed to stroke diminishes - this is because the longer the survival from the stroke, the more opportunity the patient has to die of some other cause. Across the time period from 01.01.2007 to 31.12.2011, 53% of all deaths in the Project Set of stroke patients were recorded as due to causes other than stroke and 47% were recorded as stroke.

Figure 5.5: ONS recorded Cause of Death for Project Set members, 2007-2011



5.17 The causes of death of those people admitted with a stroke emergency in 2007-09 and who had died before 31.01.2011 are shown in Table 5.3, below. The distribution of causes of death illustrates that how we select the set of patients for a Project affects the outcome. The Project used an ‘emergency admission’-focused method. Had we started by selecting patients who died from a stroke in the ONS Mortality data, around half of the patients admitted with a stroke and who later died would have been excluded; for the remainder of the group, stroke would have been listed as a co-morbidity for a different cause of death.

Table 5.3: Cause of death for people admitted with a stroke emergency 2007-09 and who had died by 31.12.2011

Cause of death	Deaths	
	n	%
Stroke	3,991	48
Other Circulatory disease codes	1,536	18
Cancer	866	10
Respiratory conditions	637	8
Mental health	262	3
Digestive system	180	2
External causes	169	2
Other causes	169	2
Genitourinary system	156	2
Endocrine, nutritional and metabolic disorders	128	2
Diseases of the Nervous system	127	2
Certain infectious and parasitic diseases.	111	1
Total	**8,332**	**100**

Source: ONS annual mortality data linked in SAIL

5.18 The examination of data so far does not reveal any simple patterns in the data that might suggest a common patient pathway shared by groups of patients. To try to establish if such patterns existed, a more specific subgroup was targeted. The young men (aged under 40 years) who were admitted for stroke were chosen. This group was chosen in order to limit the number of co-morbidities likely to be coded against them (much co-morbidity is age related); there are also more young men than young women in the set of stroke patients, so sufficient numbers existed to attempt to establish similarities in care. The analysis for young male stroke patients is reported in Chapter 6.

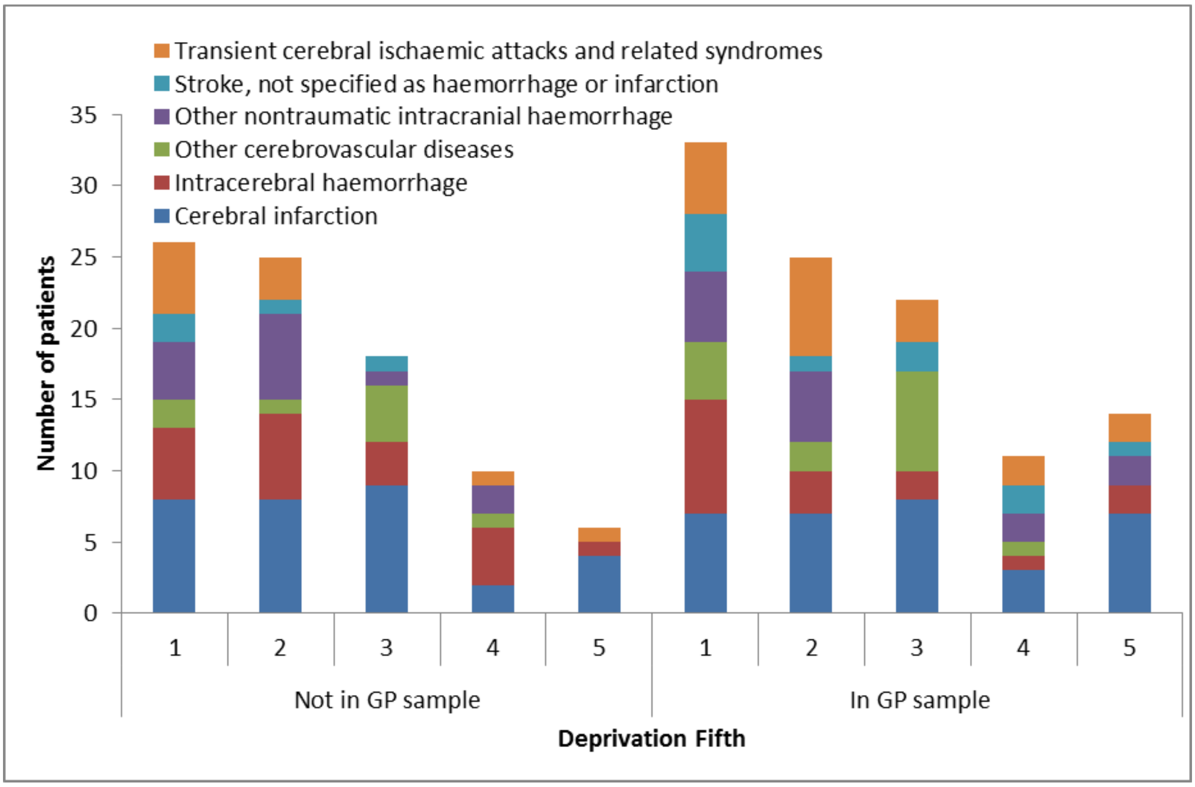
6 An in-depth look at a single subgroup of stroke patients – males aged 0-39 years

6.1 In order to test the value of the indicators in developing patient pathways, a single subgroup of the set of stoke patients was chosen. There were 192 males aged 0-39 years in the Project Set of stroke patients; 10 were babies and a further 24 were aged between 2 and 24 years. Inspection of the records revealed that 17 of these patients survived for 5 days or less, 2 survived for around a month and the rest survived for at least a year. In total, 27 (14%) had died in the period from 01.01.2007 to 31/12/2011. As noted in Chapter 1, above, the reader should bear in mind that the Project was designed to demonstrate the usefulness of linked data in allowing patient pathways to be developed; findings that relate to health service use are therefore presented as steps along the way to achieving that objective, not to serve as a report on the health of the population of Wales.

6.2 The deaths within the first week were all haemorrhagic stroke emergency admissions. Beyond this, there was no distinct pattern in terms of survival by stroke type.

6.3 Figure 6.1, below, shows the 197 young male stroke cases who were resident in Wales by deprivation fifth and stroke type, split according to whether GP data was available (107 individuals, 54%) or not (90 individuals, 46%). This indicates that there was quite a different profile of cases in the GP sample group. The GP data in SAIL is not evenly distributed geographically and this data suggests that there is an under representation of the second least deprived fifth and an over representation of the most deprived fifth of deprivation in young men in the GP practice cohort. However, these fluctuations may be explained by the small numbers involved.

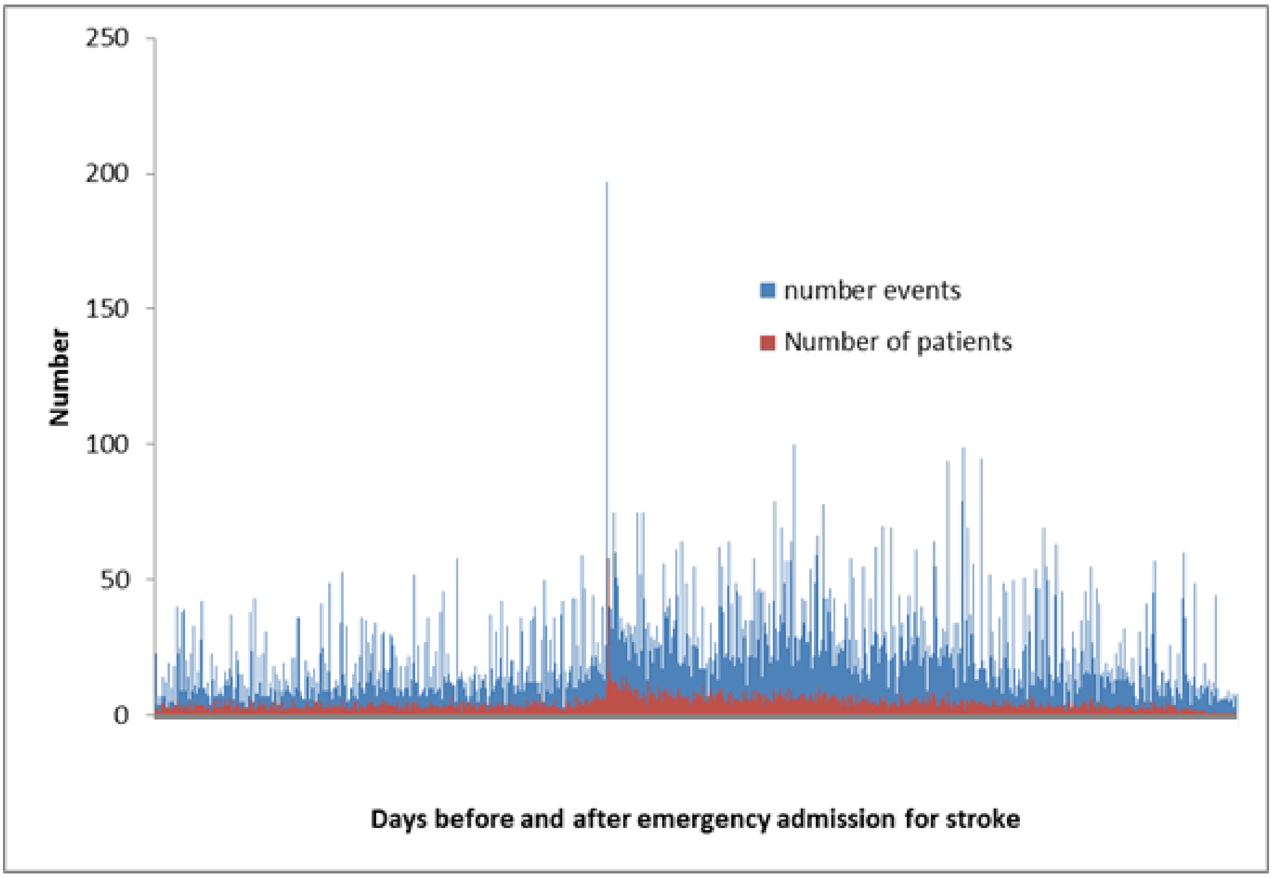
Figure 6.1 Distribution of men aged 0-39 years by deprivation, stroke type and whether detailed GP event data was available (1=most deprived)



6.4 The level of GP practice recording before and after the FEAS can be visualised by the alignment of cases. Counting the date of each person's FEAS as day zero, and labelling each day relative to day zero, we can summarise daily recording levels for the subgroup of patients leading up to and following the FEAS date, even though in reality the patients had their FEAS events on different days.

6.5 Of the 107 young male stroke patients for whom primary care event data was available, a subgroup of 89 individuals had interacted with primary care in the four years before their FEAS. For this group of 89 individuals, Figure 6.2, below, shows the number of GP records per day for the four years before and after the FEAS. The graph shows a steady underlying activity recording rate leading up to the FEAS with a large spike in data on day zero. As we would expect, following the stroke event the activity recording continues at a higher rate than prior to the event. The spike of recording on day zero is worthy of note, as it indicates that some of these men interacted with primary care while others went straight to a hospital of A&E department; in fact, on inspection, a small number of complicated patients generated the bulk of these records.

Figure 6.2 Frequency of recording at GP practice before, on and following FEAS date, for male stroke patients aged 0-39 years



6.6 Of the 107 young, male stroke patients for whom Primary care event data was available, only 58 individuals had any interaction with primary care on the date of their FEAS. These 58 individuals generated 188 records, made up of 116 different Read codes. These codes group into 'type of first contact' (e.g. A&E, telephone), stroke diagnoses and test results e.g. cholesterol level and blood pressure. Some of the records are administrative e.g. recording letters received from specialists. It is clear that some of this recording is retrospective, with the secondary care provider giving details to the GP of the investigations and subsequent actions taken in secondary care.

6.7 This provides some insight into how a more general indicator of 'level of interaction with primary care' should be defined, as some of the activity is merely administrative in nature or is being generated from outside primary care. In determining an overall GP activity level, leaving out administrative and test result codes may therefore produce a more realistic indicator.

6.8 Figure 6.2, above, does demonstrate some interaction with primary care during the period leading up to the FEAS. Analysis using the indicators developed for the Project (see Chapter 4) revealed that 75% of the young men for whom GP records were available had over two years of primary care event records but that only very small numbers (1 or 2 for each) had a record of smoking status, BMI, blood pressure, blood tests or statin or anticoagulant prescribing.

6.9 The result of this analysis showed no pattern of pre-stroke interactions with primary or secondary care for young male stroke patients. Grouping by survival and stroke type produces very small groups of patients.

7 Diabetes as a co-morbidity to stroke

7.1 As part of the development process for the project it was necessary to work out ways to introduce co-morbidities into the analysis. As shown in Table 5.2, above, the top four co-morbidities for stroke patients prior to their FEAS were either other cardiovascular conditions or a ‘Personal history of certain other diseases’. The next most common co-morbidity was ‘non-insulin dependent diabetes’, accounting for 2% of co-morbidity recording. On this basis and in discussion with policy colleagues, diabetes was chosen for further exploration as a co-morbidity. This section explores what we can establish about the effect of diabetes as a co-morbidity on the pathway and outcome for the stroke patient.

7.2 Both Insulin and non Insulin dependent diabetes were included. The 18,744 stroke patients included 3,410 (18%) with diabetes. Welsh Health Survey results suggest that the prevalence of diabetes in the adult population of Wales[14] is 6%, so the prevalence of diabetes among stroke patients was three times higher than in the general population. This is slightly higher but generally in line with the well-known increased risk of stroke in diabetics[15]. The numbers and percentages of stroke patients in the diabetic stroke group is shown in Table 7.1, below, demonstrating that a greater proportion of male than female stroke patients were diabetics. The average age of male stoke patients was 72 years for diabetics, and 70 for non-diabetics, whereas the average age of female stroke patients is 76 years in diabetics and 77 years in non-diabetics. In total, 57% of the diabetic women who had strokes had died by the end of 2011, whereas 51% of the diabetic men who had strokes had died - this may be due to women being older than men when having a stroke.

[14] http://wales.gov.uk/docs/statistics/2011/110913healthsurvey10en.pdf

[15] http://link.springer.com/article/10.1007/s00125-006-0493-z#page- Study of stroke risk in type 2 diabetes using General Practice Research Database in the UK.

Table 7.1 Number and percentage of stroke patients by whether had diabetes and gender

	Unit	Men	Women	People
Diabetes	N	1801	1609	3410
No diabetes	N	7161	8173	15334
Total	N	8962	9782	18744
Diabetes	%	20	16	18
No diabetes	%	80	84	82

7.3 The complete set of stroke patients (including those with no primary care record) was examined to determine which patients had a coding in secondary care indicating they were diabetic. The relationship between the type of stroke and the type of diabetes is shown in Table 7.2, below.

Table 7.2: Number of patients by Stroke and Diabetes diagnosis

Stroke type	No diabetes		Insulin dependent diabetes mellitus		Non- insulin dependent diabetes mellitus		All other Diabetes Mellitus	
	No.	%	No.	%	No.	%	No.	%
Transient cerebral ischaemic attacks and related syndromes	2814	18%	50	15%	547	18%	<5	10%
Retinal vascular occlusions	11	0%	<5	0%	<5	0%	<5	0%
Intra-cerebral haemorrhage	1208	8%	25	7%	169	6%	<5	14%
Other non-traumatic intracranial haemorrhage	538	4%	8	2%	104	3%	<5	7%
Cerebral infarction	6234	41%	150	45%	1313	43%	14	48%
Stroke, not specified as haemorrhage or infarction	2969	19%	61	18%	613	20%	5	17%
Occlusion and stenosis of pre-cerebral arteries, not resulting in cerebral infarction	164	1%	12	4%	34	1%	<5	0%
Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction	17	0%	<5	0%	<5	0%	<5	0%
Other cerebrovascular diseases	1347	9%	29	9%	260	9%	<5	3%
Sequelae of cerebrovascular disease	32	0%	<5	0%	<5	0%	<5	0%
Total	15334	100%	335	100%	3046	100%	29	100%

7.4 There was no relationship between diabetes and type of stroke. The only noteworthy differences were 4 percentage point fewer transient ischaemic attacks and 4 percentage point more cerebral infarctions in insulin dependent diabetics compared with non-diabetics.

7.5 Further comparison of the diabetic and non diabetic stroke patients indicated that in the four years leading up to and including the FEAS, male diabetics had spent significantly longer in hospital than non-diabetics (table not shown). The average length of stay in non-diabetic male stroke patients was 41 days (95% C.I. 40.2 to 41.8) whereas the average length of stay for diabetic stroke patients was 57 days (95% C.I. 47.8 – 66.2), on average a stay of 16 days longer per patient. The additional stay was highest in Type 1 diabetic (insulin dependent) men who had an average length of stay prior to and including their FEAS of 92 days (95% C.I. 50.4 to 133.6).

7.6 The picture is different for female patients. As reported earlier, the overall average length of stay in hospital prior to and during FEAS was significantly longer for women (61 days, 95% C.I. 56.2 to 65.8) than for men (44 days, 95% C.I. 39.6 to 48.8). However the average length of stay in non-diabetic and diabetic women was very similar (62 and 60 days respectively). The average length of stay for women with Type 1 diabetes was 64 days. The increased length of stay seen in Type 1 diabetic men is not seen in women with Type 1 diabetes.

7.7 These observations about lengths of stay may be in some way explained by the relatively older female stroke population. However, it is also possible that diabetes management may be better in female than in male insulin dependent patients. This finding coupled with the distinct age difference between the male and female stroke patients, suggests that analysing the health service experiences of stroke patients should ideally be done separately for men and women. Within the limited scope of the Project it was not possible to repeat the analysis by gender but further potential work is described in Appendix 5.

7.8 As a result of the analysis completed for the Project, we can conclude that it is possible to identify different pathways for patients with one or more co-morbidities, or different co-morbidity combinations, but that every patient characteristic e.g. age group and gender, may require quite different sets of pathways. Within the limited scope of a demonstration project, it has not been possible to examine the interaction of more than a single co-morbidity, but the analysis developed so far is close to achieving that aim and has provided some groundwork on which future projects could build. However, it should be noted that the amount of time required to robustly define each condition and to work out how to model its interaction with other conditions should not be underestimated.

7.9 The whole area of patient pathway definition and modelling patient flows is an area of significant international focus, with the aim of reducing unnecessary hospital admissions through prevention, and maximising both the efficiency and effectiveness of hospital services[16]. This project has contributed to this growing field of evidence and has made suggestions for ways further projects might build on this work (see Appendix 5). Examining the summary data in Chapters 5, 6 and 7 suggests that some clearly defined rules would need to be established to determine how to split sets of patients into those with similar patient pathways, and that differences in pathways will not be immediately obvious from undertaking simple frequency or crosstabulation analysis of the data. In order to establish whether this population of patients could be divided into specific groups based on the chosen indicators, a cluster analysis was carried out.

[16] For example: Vascular disease in women: comparison of diagnoses in hospital episode statistics and general practice records in England F Lucy Wright*, Jane Green, Dexter Canoy, Benjamin J Cairns, Angela Balkwill, Valerie Beral: http://www.biomedcentral.com/1471-2288/12/161

8 Cluster analysis of stroke patients

8.1 In this section, the results of the experimental cluster analysis are presented. As noted in Chapter 4, the analysis was run specifying solutions with between 7 and 10 clusters. By examining the solutions, it was clear that the 10-cluster solution demonstrated the greatest homogeneity within the clusters. However, it was also clear that the clusters retained significant variation so although within the limited scope of the demonstration project it was not possible to re-run the analysis to include greater numbers of clusters, a future project, as suggested in Chapter 4, could be completed to use a different form of clustering to identify the ideal number of clusters before completing the analysis using k-means.

8.2 As noted above, it was important to analyse the interaction patients had with both primary and secondary care. The cluster analysis could therefore only be completed where primary care event data existed for the patient. The total set of stroke patients for whom primary care event data was available included 8,781 patients, but we chose to exclude patients for whom there was insufficient longitudinal data or who had retinal vascular occlusions, as described in Section 4.31. This left a total of 6,921 cases to be clustered.

8.3 The characteristics of each cluster for the 10-cluster solution are summarised in Table 8.1, below. The summary descriptions are based on the characteristics where the cluster was demonstrated to have significantly different membership compared with the average; for example, a cluster summary will describe the cluster as being ‘male’ if the cluster contains significantly more men than average. The more detailed cluster descriptions presented in the right-hand column summarise the individual items where differences were significant; where items are in bold, this means there was a particularly significant difference when compared with the average[17].

8.4 The clusters range in size from 1774 to 78 stroke patients. Deprivation score, area classification, and anticoagulant prescribing were determined as having

[17] Significant differences are based on adjusted standardised residuals of greater than 1.96; bold differences are based on adjusted standardised residuals of greater than 10.

no significant effect on the clustering of patients when clustering was performed. These items have therefore been excluded from the summaries below.

Table 8.1: Summary of 10-Cluster Solution: Stroke patients

Cluster summary description **(based on average for cluster)**	**Cluster size**	**Cluster contained significantly … than average**
High proportion of younger adults with small amount of previous (non-emergency) hospitalisation and primary care input but only very recent statin prescription having an event from which they survive for more than 2 years	1,774	• more aged 40 to 49 years • more with the stroke types: occlusions, 'strokes' and 'other CVD' • **more who were still alive after 2 years** • more with zero or one previous emergency admission • more with previous total hospital stay of between 8 days and six months, **particularly between 15 and 28 days** • **more with no diabetes** • more with 1 to 110 previous GP interaction dates, **particularly between 1 and 16** • more with fewer than 2 previous statin prescription dates • more with fewer than 8, **particularly 4 or less** BP monitoring events • **fewer with weight or BMI recorded** • more with fewer than 10 blood tests of some kind before, **particularly with none** • **more never asked** before about smoking
Younger adults with diabetes and some previous emergency admissions but relatively little primary care input or monitoring having a less serious event from which they survive for more than two years	391	• more aged under 60 years and fewer aged 70-79 years • **more with the stroke types: TIA** and Haematoma • more still alive after 2 years • more with two or more previous emergency admissions • more with between one and six months total previous hospital stay • more with Insulin dependent diabetes • **more with 16 or fewer** previous GP interaction dates • **more with fewer than 2** previous statin prescription dates • **more with no** BP monitoring events • **fewer with a weight or BMI recorded** • **more with no** blood test of some kind before • **more never asked** before about smoking
Men aged 60 to 79 years with complex health issues having a relatively minor, long-survived event	722	• more men • more aged 60 to 79 years • **more TIAs** and fewer of all other stroke types • **more still alive after 2 years** • more with no previous emergency admissions • **more with 7 days or less** total previous hospital stay • **more with Non-insulin dependent diabetes** • more with 66 or more previous GP interaction dates, **particularly with 111 to 179** • more with between 2 and 182 previous statin prescription dates, **particularly between 29 and 182** • more with 9 or more BP monitoring events, **particularly 17 or more** • **more with weight or BMI recorded** • **more with 21 or more** blood tests of some kind before • more asked 6 or more times about smoking, **particularly between 6 and 10 times**

Cluster summary description **(based on average for cluster)**	**Cluster size**	**Cluster contained significantly … than average**
Men aged under 70 with some previous hospitalisation and history of smoking and primary care monitoring for weight/BMI and BP having stroke or other cerebrovascular disease and surviving more than 2 years	1,126	• more men • more aged under 70 years and fewer aged 80-89 years • more 'stroke' and other CVD • **more still alive after 2 years** • **more with no** previous emergency admissions • more with fewer than 28 days total previous hospital stay, **particularly between two and seven days** • **more with no diabetes** • more with between 1 and 65 previous GP interaction dates, **particularly between 17 and 65** • **more with fewer than 2** previous statin prescription dates • more with between 9 and 16 BP monitoring events • more with 20 or fewer blood tests of some kind before • more asked between 1 and 5 times about smoking
People aged 50 to 69 years with diabetes and long history of primary care advice on smoking but no monitoring for bloods, BP or weight having an event from which they survive for more than 2 years	78	• more aged 50-69 years and fewer aged 80-89 years • fewer haematomas • more still alive after 2 years • more with no previous emergency admissions • more with between one and seven days total previous hospital stay • **more with Non-insulin dependent and ‘Unspecified diabetes’** • more with 16 or fewer previous GP interaction dates • more with fewer than 2 previous statin prescription dates • more with no BP monitoring events • fewer with weight or BMI recorded • more with no blood tests of some kind before • more never asked about smoking
People aged 80-89 years with primary care advice on smoking and monitoring for BP, bloods and weight/BMI with diabetes and significant primary and secondary care input having a stroke or other cerebrovascular disease and surviving between six months and two years	476	• more aged 80-89 years and fewer aged under 39, 50-59 and 90+ years • more with 'stroke' and other CVD • more dying at between six months and two years • more with 2 or more previous emergency admissions, **particularly 3 or more** • more with between one month and two years total previous hospital stay, **particularly between one and six months** • more with insulin dependent diabetes, **particularly more with Non-insulin dependent diabetes** • **more with 111 or more** previous GP interaction dates • more with between 2 and 364 previous statin prescription dates, **particularly between 29 and 182** • more with 9 or more BP monitoring events, **particularly with 17 or more** • **more with weight or BMI recorded** • **more with 21 or more** blood tests of some kind before • more asked about smoking 6 or more times, **particularly between 6 and 10 times**

Cluster summary description **(based on average for cluster)**	**Cluster size**	**Cluster contained significantly … than average**
Women aged 80+ years with significant primary and secondary care input and diabetes as a co-morbidity having catastrophic event with very short survival	1,173	• more women • **more aged 80+ years and fewer aged under 80 years** • more haematomas and occlusions • fewer still alive after 2 years, **particularly more dying between 7 days and 6 months** • more with one, two or three previous emergency admissions • more with 15 days or more total previous hospital stay, **particularly between 15 days and six months** • more with 'Other specified diabetes' • more with 66 or more previous GP interaction dates, p**articularly with 111 to 179** • more with between 2 and 28 previous statin prescription dates • more with 9 or more BP monitoring events, **particularly with 13 or more** • **more with weight or BMI recorded** • **more with 21 or more** blood tests of some kind before • more asked at least once about smoking
Oldest women with significant primary and secondary care input having relatively minor event (this cluster may profit from splitting, since it contains significantly more than average of two separate survival groups and two distinct diabetes groups)	365	• more women • more aged 90+ years and fewer aged 60-69 years • more TIAs • more surviving between one and six months and between 1 and 2 years • **more with 3 or more** previous emergency admissions • more with between one month and two years total previous hospital stay, **particularly between one and six months** • more with no diabetes and more with Insulin dependent diabetes • more with 66 or more previous GP interaction dates, **particularly with 180 or more** • more with fewer than 3 previous statin prescription dates • more with between 1 and 8 BP monitoring events • more with 11 or more blood tests of some kind before • more asked between 1 and 5 times about smoking
Oldest individuals, relatively un-monitored & untreated having catastrophic event with very short survival	620	• more aged 90+ years • more TIAs and **Haematomas** • **more surviving less than one month** • **more with no** previous emergency admissions • **more with 7 days or less** total previous hospital stay • more with no diabetes • more with 17 to 110 previous GP interaction dates • more with fewer than 2 previous statin prescription dates • more with between 1 and 8 BP monitoring events, **particularly between 1 and 4** • more with weight or BMI recorded • more with 11 or more blood tests of some kind before • more asked between 1 and 5 times about smoking

Cluster summary description **(based on average for cluster)**	**Cluster size**	**Cluster contained significantly … than average**
Oldest individuals, almost completely un-monitored & untreated suffering single, serious event with relatively long survival	203	• more aged 90+ years • more haematomas and 'strokes' • more surviving less than a year, **particularly less than one month** • more with between 3 and 5 previous emergency admissions • more with between 15 and 28 days total previous hospital stay • more with 16 or fewer previous GP interaction dates, **particularly with none** • more with fewer than 3 previous statin prescription dates • **more with no** BP monitoring events • **fewer with weight or BMI recorded** • **more with no** blood tests of some kind before • **more never asked** about smoking

8.5 As we would expect from the previous observations of age and gender differences, some of the resulting clusters are significantly different to the average according to age group and gender - the younger patients tend to appear in distinct clusters and some clusters are strongly gender-related.

8.6 Some of the indicators that indicate patient monitoring with primary care suggest a distinction between those patients who were relatively highly engaged with primary care and those who were not. Some of the clusters contain groups of patients with relatively minor stroke events and no diabetes but relatively high levels of monitoring in primary care and interactions with both primary and secondary care, suggesting that they contain patients with other conditions not included in the analysis. This may be due to the presence within the cluster of patients with just the kinds of multiple chronic conditions the project was originally designed to identify but further work would be required to explore this possibility further.

8.7 This clustering should be considered indicative rather than a final solution. There are several ways that the clusters might be modified in order to produce a different result. The key improvements would be:

- to re-run the clustering using a different method to identify the ideal number of clusters, as described in Chapter 4; and
- to include further indicators or to categorise the existing indicators differently.

8.8 Nevertheless, the analysis has shown that meaningful clusters of patients with similar experiences of health service use prior to stroke can be identified. Further work might be done to refine the solution for stroke patients. However, the principle having been established that this kind of analysis can identify meaningful groupings, it could be used to undertake additional analysis for further main conditions and to examine the effect of including more than one co-morbidity in the cluster analysis.

9 Discussion and conclusions

9.1 By using data linked across several health datasets, this project attempted to identify groups of patients who had similar patient pathways. There was particular interest in the impact of co-morbidities.

9.2 The project demonstrated a methodology that could be used to identify a disease-specific cohort of anonymised patients presenting as emergency admissions to hospital for a specific cause, and to link these patients across a number of datasets to provide a detailed picture of their interactions with primary and secondary healthcare prior to a 'benchmarking' event. This has been developed for a stroke emergency admissions group but could be used for any disease group. A second group of people who had emergency admissions for diabetes was generated to test this but within the limited scope of a demonstration project, it was only possible to use diabetes as a co-morbidity in the analysis of stroke.

9.3 The methodology includes aligning individual timelines of primary care events according to a specific 'benchmark' event, in this case the first emergency admission for a stroke. A series of indicator variables were created to that could be used to compare primary and secondary care events before and after the 'benchmark' event. However, within the limited scope of a demonstration project, it was only possible to examine indicators describing the pre-stroke interaction of stoke patients with primary and secondary care.

9.4 The information generated was scrutinised in a number of ways to establish whether those experiencing emergency stroke admissions could be subdivided into specific groups of patients with different pathways. In particular, there was interest in identifying the impact on the pathway where the patient was diagnosed with more than one chronic condition. The Project was able to examine the impact on patient pathways of the single co-morbidity of diabetes. The number of combinations of different co-morbidities made further definition of specific care pathways impossible within the limited scope of a demonstration project.

9.5 Some limitations in the methodology have been identified; these are discussed in detail in Chapter 4. Because the analysis of linked administrative data is a new field and the project was therefore both developmental and experimental, a large number of analytical decisions and assumptions had to be made in order to establish the methodology. The outputs presented in the Report make it clear that a deeper understanding of the data is required. Multi-disciplinary expert input would be a necessary next step to refine and review these decisions before any weight could be given to the findings or the process taken to another stage but these steps were not possible within the limited time available for the demonstration project.

9.6 Nevertheless, the project has demonstrated that there is potential to define a number of specific groups of patients who experience a similar pathway prior to a stroke emergency. A single co-morbidity was built into the cluster analysis but additional work would be required to define further co-morbidity combinations and to examine their impact.

9.7 The key findings with regard to the stroke pathway are:

- That patient pathways vary significantly by both age group and gender.
- That patient pathways tend to be split between groups with high and low levels of interaction with primary and secondary care, with some groups having almost no interaction before the event.
- The clustering identifies some groups of patients with relatively minor stroke events and no diabetes but relatively high levels of monitoring in primary care and high levels of interaction with both primary and secondary care. This is an emerging method that could be used to determine which groups of patients to focus on for further analysis. Further work would be required to choose, refine and incorporate specific combinations of co-morbidities before a tool could be developed to identify sets of typical patient pathways for patients with multiple chronic conditions.

9.8 Having used the cluster analysis to identify groups of stroke patients with similar pre-stroke interactions with primary and secondary care, further work would be required to develop better methods of visualisation / presentation to allow such groups to be presented in ways that are both engaging and useful to the policy customer and to practitioners.

9.9 The complex nature of the data has meant that a significant amount of project time went into developing an understanding of this data e.g. how to manipulate and recode it into suitable forms to use as indicators. Despite repeatedly narrowing the scope of the project to focus on more and more specific disease groups, the process has remained so time consuming that consideration of other datasets such as Outpatients and Accident and Emergency datasets has not been possible.

9.10 In order to narrow the focus of the Project, it has focussed on patient interactions with primary and secondary care *before* the stroke event. It would also be possible for future projects to analyse the *post stroke* experience.

Next steps

9.11 As the fellowship time has come to an end, the question arises as to whether and how this work can be taken forward from here. I would estimate it would require a full time developer a year to get to the point of having a tool to e.g. compare patient treatments for different outcomes. The demonstration project has proceeded far enough to enable the informed writing of a specification for such a tool, but such developments require funding, and before seeking funding, a review of what overlapping projects and initiatives are already in place to deliver something similar.

9.12 One way forward would be to identify a research project requiring patient pathways to be identified for a specific condition, either with or without a specific treatment in mind. The SQL code developed for the demonstration project could be adapted to complete the analysis for such projects. In 2012, SAIL was successful in bidding to lead the Centre for the Improvement of Population Health through e-Records Research (CIPHER), one of the four major centres that will make up the Farr Institute, a MRC-funded UK health informatics research institute. The creation of the Farr Institute will present opportunities to identify projects that can build on the methods this Project has developed

9.13 Some of the code development completed for this Project will be fed into the process underway at SAIL to make data more readily available for research.

The code removes data duplication and performs a number of recoding processes that could be adopted for standard use. Part of the work of the Centre for Improvement of Population Health through e-Records Research (CIPHER), the MRC e-Health Improvement Research Centre based at SAIL, is about developing standardised methodologies for defining and refining administrative data in this way.

9.14 On a personal note, this process has been incredibly educational, in many ways. The Author has learned a lot about how long such projects can take and how inadequate our understanding is of some of the datasets held by SAIL. One particular lesson is the significance of missing information when attempting to analyse linked datasets - in primary care the recording tends to be of ‘abnormal’ signs and symptoms, so the majority of some fields may be blank and we are left wondering whether the absence of the ‘abnormal’ means that we can assume normality.

9.15 Throughout the Project, we have worked closely with NISCHR, the part of WG that funds SAIL, to feed back about areas where SAIL might improve its service and this has influenced NISCHR’s funding of SAIL-related activities. The lessons learned from the suite of demonstration projects will feed into the further development of SAIL, and into the work programmes of both CIPHER and the new Wales Administrative Data Research Centre. Not only will the work completed for these projects inform specific future projects but they will allow analysts and customers to be given a better idea of how long similar projects may take.

Appendix 1: Membership of the Information Governance Review Panel

N.B. Membership provided is for 10.06.2013

The IGRP provides independent advice on Information Governance and reviews all proposals to use SAIL data to ensure that they are appropriate and in the public interest. The current panel comprises of

Organisation	Member
British Medical Association	Dr Tony Calland
National Research Ethics Service	Corrine Scott
Public Health Wales	Dr Judith Greenacre
NHS Wales Informatics Service	Martin Murphy Darren Lloyd
SAIL Consumer Panel	Dr Neil McKenzie Dot Williams

Appendix 2: HIRU Application form (IGRP application)

chiral Centre for Health Information, Research and Evaluation
Canolfan Iechyd Hysbysrwydd, Archwiliad ac Ymchwil

For HIRU use
Ref. No.

Centre for Health Information Research and Evaluation (CHIRAL)

College of Medicine

Swansea University

Health Information Research Unit (HIRU)

HIRU Enquiry form

Template review chronology

Version no.	Effective date	Reason for change
1.0	29/11/07	N/A
2.0	1/5/08	Establishment of CRS necessitating changes to content and layout
3.0	14/10/09	Recommendations of IGRP
3.1	05/04/11	Annual review

The following form has been designed to collect the information needed from individuals and organisations interested in collaborating with HIRU on work involving the SAIL databank. The information you provide will facilitate consideration of your enquiry. Please complete sections and A & B and provide additional documents as requested.

SECTION A

1a. Contact details of project lead:

Name:

Job title:

Organisation:

Address:

Tel:

Fax:

Email:

1b. The project lead will be the only person accessing the data:

Yes [□] No [☒]

1c. Please provide contact details of the person(s) who will be accessing the data (apart from the project lead):

Name:

Job title:

Organisation:

Address:

Tel:

Fax:

Email:

2. Does your proposed work with HIRU constitute:

Part of a larger project?

If yes, please complete all questions

The entire project?

If yes, please complete all questions except 3a, 5a and 7a

3a. Full title of the *main* project:

3b. Full title of the (part of the) project involving HIRU (if different):

4a. Who is commissioning the project (if relevant)**?**

4b. Why is the project being done?

5a. Aim of the *main* project, including anticipated outcomes:

5b. Aim of the (part of the) project involving HIRU, including anticipated outcomes (if different):

Please include a copy of the protocol/plan for the proposed work with HIRU, including the contact details of any co-applicants when you return your completed form.

6. Lay summary of the project involving HIRU: (approximately 150 words)

7. Please list the relevant permissions you have obtained or that are being sought:

	Obtained	*Being sought*	*Not required*
Research ethics	[☐]	[☐]	[☒]
Independent peer review	[☐]	[☐]	[☒]
Permission from data-holding organisation to use their datasets	[☐]	[☐]	[☒]

Please state the name of the organisation/committee that is being applied to, or that has given approval, as applicable:

Research ethics:

Peer review:

Data organisation permission:

If you have ticked 'not required' please specify the reasons.

Please note that it is the responsibility of the project lead to ensure that the relevant permissions are obtained.

8a. At what stage is the main project?

Protocol/plan being developed [☒]

Protocol/plan in place but project not commenced [□]

Project underway [□]

If underway, what was the start date of the main project (dd/mm/yy)?

8b. Please indicate a prospective start date for the (part of the) project involving HIRU:

(dd/mm/yy)

8c. Over what period do you anticipate you will require the assistance of HIRU?

Start and end dates in dd/mm/yy: [ ] to [ ]

9a. What data do you require for the proposed work with HIRU?

Please list:

The datasets you require information from

The types of variable you need

The datasets that will need to be linked

9b. Will you also be providing other datasets to be incorporated into the SAIL databank?

Yes [□] No [☒]

If yes, please specify:

9c. Please provide an outline of your analysis plan including the anticipated outcomes

9d. Are the results/methods developed likely to have other potential applications?

Yes [☒] No [□]

If yes, please specify:

10a. Please indicate your plans for publishing the results of your project, e.g. target journal or intended recipients of report:

10b. What are the potentially sensitive issues that need to be taken into account when publicising the findings of the project?

Please outline the issues and your proposed solutions:

Appendix 3: Overview of datasets used in this project

As noted in Chapter 1, the reader should bear in mind that the Project was designed to demonstrate the usefulness of linked data in allowing patient pathways to be developed; the report presents analyses as steps along the way to achieving that objective, not to provide the definitive statistical outputs for the data sets described below.

The text and diagrams provided in this Appendix mention two concepts that need definition – ‘Entity Relationship Diagrams’ and the acronym ‘MAI’:

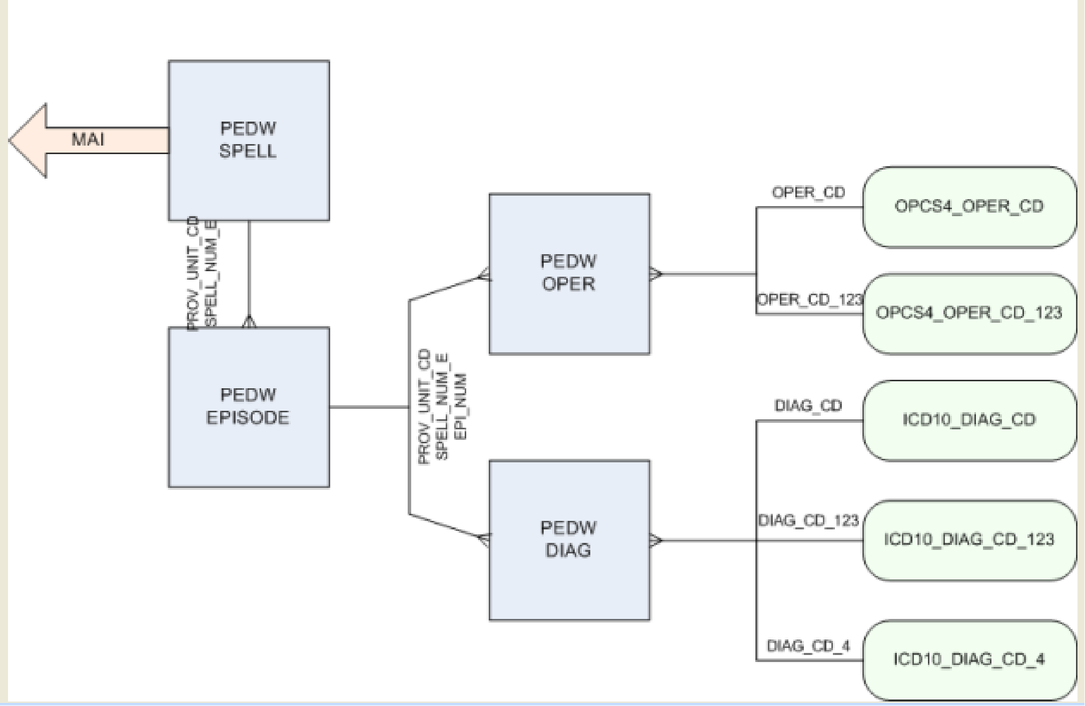
- Entity relationship modelling is a Software Engineering concept, providing an abstract way of describing a database. SAIL comprises of a set of relational databases, i.e. data is logically separated into separate but related tables for efficiency of storage and speed of data manipulation. Some of the data in these tables ‘point to’ data in other tables – so, for example, a ‘person’ in the WDS database could point to several entries for each of the ‘addresses’ they have lived in. For the purposes of the entity relationship diagrams shown below, for example, each ‘person’ is an entity and each ‘address’ is an entity and the relationship between the ‘person’ and the ‘addresses’ would be 'has an address'. Diagrams created to represent these entities and relationships are called entity–relationship diagrams or ER diagrams.
- The SAIL entity relationship diagrams provided below all refer to the ‘MAI’. The MAI is the Master ALF Index - all the ‘person’-based databases in SAIL have a relationship with the ‘Master Database’ relating to the People of Wales.

The PEDW data structure

- The Patient Episode Database for Wales (PEDW) is the national repository for in patient and day case data. LHBs in Wales are required to download, on a monthly basis, very clearly defined and standardised data from all hospital Patient Administration Systems (PAS). These are collated by the NHS Wales Informatics Service (NWIS), who also receive details of Welsh patients treated in England through a mechanism known as the NHS switching service, and provide details of English patients treated in Wales to Hospital Episode Statistics (HES), the equivalent system in England.

- PEDW is an all-Wales database containing all finished consultant episodes of in-patient or day case care carried out in Wales, and treatments carried out on Welsh residents elsewhere in the UK. A finished consultant episode is defined as a completed ‘unit’ of care under the care of one consultant. Each episode has provision for a number of diagnosis and operative procedure codes to be recorded. In PEDW, the ICD 10 diagnostic codes are utilised. So, for example, first episodes of care containing a diagnosis in the range I00-I99 relate to episodes of cardiovascular diseases.
- ‘Finished Consultant Episodes’ (FCEs) that occur over an uninterrupted contiguous time period form a ‘spell’ of care, which are defined as the complete set of finished consultant episodes making up a contiguous period of time spent in a hospital. [Ref data dictionary] The PEDW data is organised in SAIL according to the Entity Relationship Diagram shown in Figure A3.1. The Spell table contains the ALF that allows the patient to be linked to other datasets in SAIL. The fields ‘provider unit code’ and ‘spell number’ allow linkage into the Episode table, so that all episodes can be found relating to the spells for an individual patient. Three fields, ‘Provider unit code’, ‘Spell number’ and ‘Episode number’ are used to link the episode table to the diagnosis table and the procedure tables.

Figure A3.1: Entity relationship diagram for PEDW data in SAIL

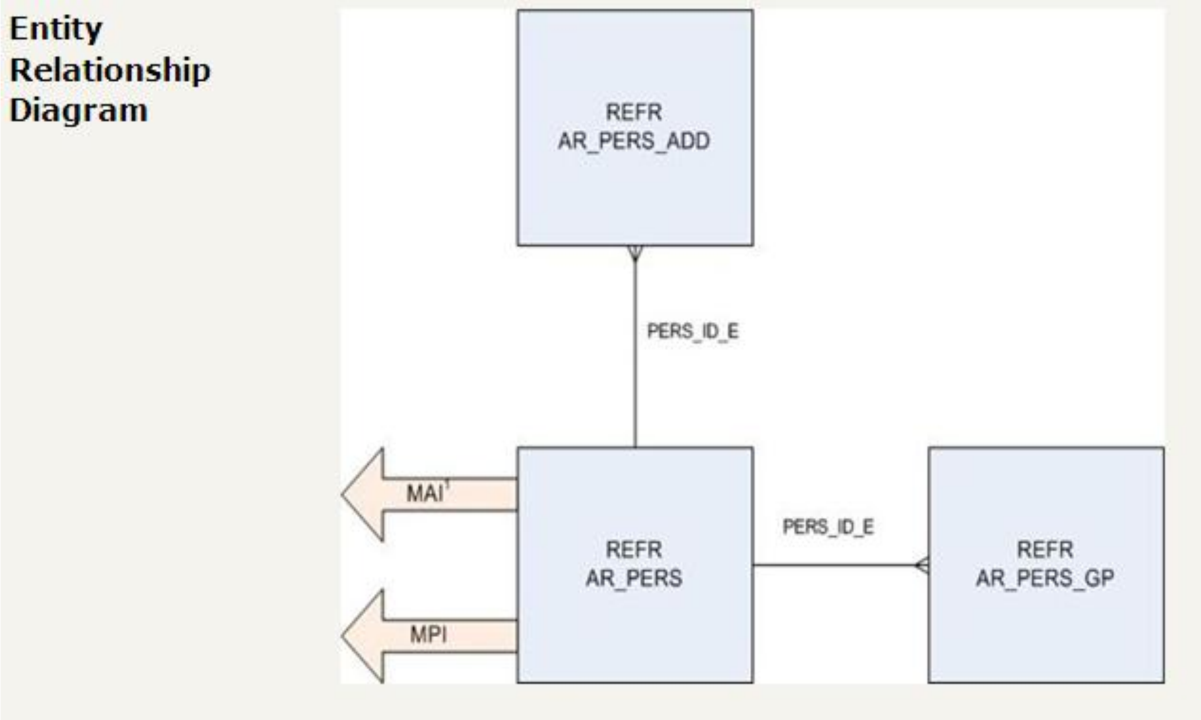


Welsh demographic service data

The Welsh Demographic Service is a dataset of administrative information about individuals in Wales that use NHS services, such as address and GP practice registration history. It replaced the NHS Wales Administrative Register (NHSAR) in 2009. This dataset contains the full registration history of the population of Wales since 1990, including house moves and changes of registration to different GP practices. This is the core data that is used in linking datasets together in SAIL. Each person's week of birth is recorded and a date of death when known.

The data is stored in three views as shown in Figure A3.2, below, with a core list of people (as ALF codes, that is) in the AR_PERS view. As people can have several addresses over time the AR_PERS_ADD view contains each change of address in a separate row. Also as each patient can change GP and hence have a number of registrations, this information is stored in a separate view AR_PERS_GP.

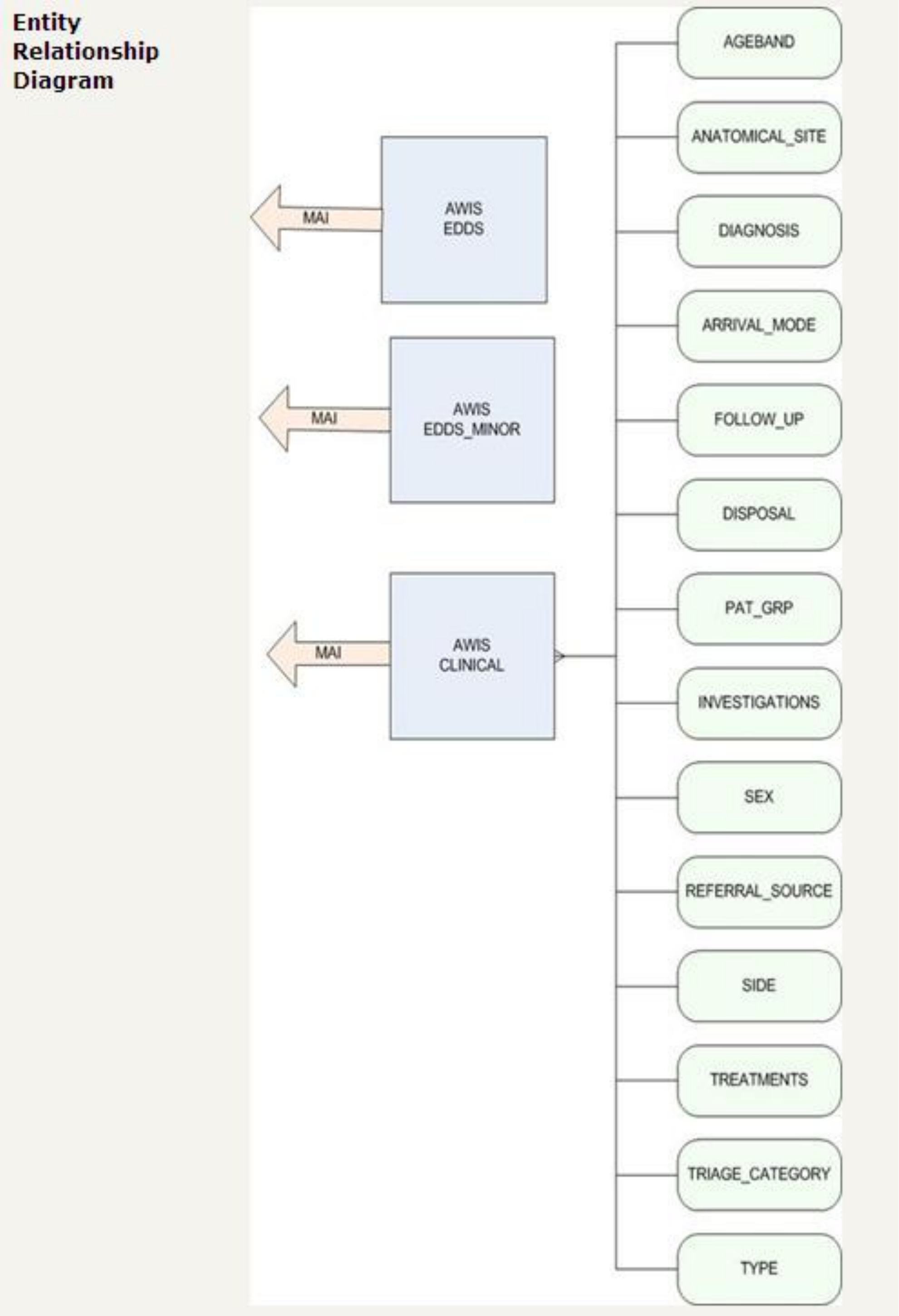
Figure A3.2 Entity relationship diagram for Welsh Demographics Service (formerly Administrative Register – AR) data in SAIL



Accident and Emergency department data

Historically data about Accident and Emergency visits was recorded in SAIL from the All Wales Injuries and Surveillance System. From 2009 this was superseded by the Emergency Department Dataset (EDDS), which contains administrative and clinical information for all NHS Wales Accident and Emergency department attendances, which now includes the (AWISS) data. The entity relationship diagram for this dataset is shown in Figure A3.3 below.

Figure A3.3 Entity relationship diagram for EDDS data in SAIL



The resulting clinical data from both major and minor A&E units is stored in AWISS clinical, with the fields depicted on the right in the above diagram. There are however two distinct linking tables to provide the means to associate these records with other SAIL datasets. As some stroke patients may have had their ‘first contact’ with either a major or a minor unit, this creates complications when selecting cases using database queries.

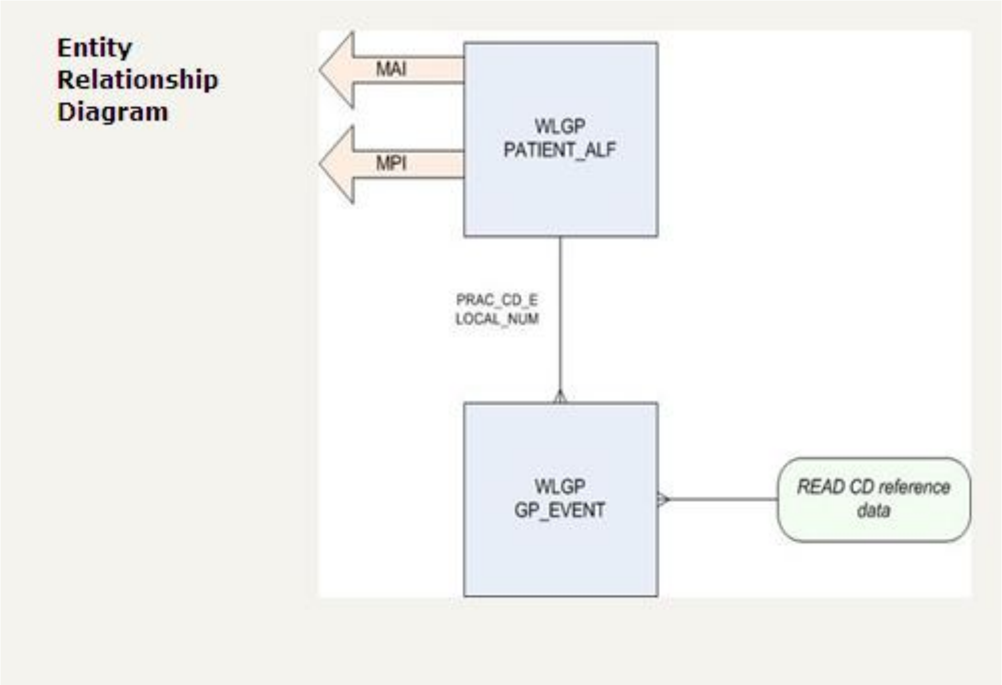
GP EVENT DATA

This is data extracted from all Welsh General Practices that have signed up to SAIL. The data is from the clinical information system the practice uses to maintain an electronic health record for each of their patients - capturing the signs, symptoms, test results, diagnoses, prescribed treatment, referrals for specialist treatment and social aspects relating to the patients home environment. The majority of the data is entered by the clinician during the patient consultation, though the data also record interaction with other members of the practice team, repeat prescribing, and some test results that are reported back from secondary care systems. The data cover the period from January 2000 to August 2012, approximately, but this varies by practices. Currently about 47% of the Welsh population is included in this dataset.

There are no standard rules for recording data within primary care clinical information systems. Therefore, each individual clinician can record information in their own way. The majority use Read Code Terminology, however, sometimes this is applied behind the scenes by the clinical system and sometimes local codes are used. Read codes are not as precise as ICD 10 or OPCS codes. Coding standards have been agreed on for conditions monitored by the QOF (Quality Outcomes Framework) returns. Since the implementation of QOF these conditions have been coded in a more consistent way.

The data format is simple, as presented in the Entity relationship diagram in Figure A3.4, below. Essentially each item of recorded information is stored in a single row in GP_EVENT, with a Read Code for the item, an event date (when the event occurred) and an optional corresponding event value. So if the event was that the patient’s blood pressure was recorded there will be a code for this event, a date when it occurred and possibly a value entered indicating what the blood pressure reading was.

Figure A3.4 Entity relationship diagram for GP event data



Within this simple data structure a lot of detailed information can be presented. Table A3.1, below, shows the diversity of possible entries for a single measurement – smoking status. This includes 49 different codes ranging from not being a smoker through various smoking types to being an ex – smoker. Over time, the same individual may be coded to several of these smoking status codes. Similar complexity is seen in many diagnostic, sign and symptom, tests and procedure coding in the GP data.

Table A3.1 Read code complexity: codes available to GPs to record smoking status

Read codes related to smoking status of patients			
Read	**Description - Ex smokers**	**Read**	**Description Current smokers**
1377	Ex trivial smoker	1372	Trivial smoker
1378	ex light smoker	1373	light smoker
1379	Ex moderate smoker	1374	Moderate smoker
137A.	Ex heavy smoker	1375	heavy smoker
137B.	ex very heavy smoker	1376	Very heavy smoker
137F.	Ex-smoker amount unknown	137..	tobacco consumption
137j.	ex cigarette smoker	137a.	Pipe tobacco consumption
137K.	Stopped smoking	137b.	Ready to stop smoking
137K0	Recently stopped smoking	137c.	Thinking about stopping smoking
137L.	current non-smoker	137C.	Keeps trying to stop smoking
137l.	ex roll-up cigarette smoker	137d.	Not interested in stopping smoking
137N.	ex pipe smoker	137D.	Admitted tobacco consumption untrue

137O.	ex cigar smoker
137S.	ex-smoker
137T.	Date ceased smoking

Read	Description - Non-smokers
1371	never smoked
137I.	Passive smoker
137U.	Not a passive smoker

137E.	Tobacco consumption unknown
137e.	Smoking restarted
137f.	Reason for restarting smoking
137G.	Trying to give up
137g.	Cigarette pack years
137H.	Pipe smoker
137h.	Minutes from waking to first tobacco
137J.	Cigar smoker
137k.	Refusal to give smoking status
137M.	Rolls own cigarettes
137m.	Failed attempt to stop smoking
137P.	cigarette smoker
137Q.	Smoking started
137R.	current smoker
137V.	Smoking reduced
137W	Chews tobacco
137X.	Cigarette consumption
137Y.	Cigar consumption
137Z.	tobacco consumption NOS

Appendix 4: Details of data preparation for Cluster Analysis

In order to perform cluster analysis, the variables were clustered into categories based on inspection of the distribution of values in the data.

SURVIVAL in days was recoded as

(0=1) (1 thru 4 = 2) (5 thru 9 = 3) (10 thru 14 = 4) (15 thru 19 = 5) (20 thru 24 = 6) (25 thru 29 = 7) (30 THRU 59 = 8) (60 thru 89 = 9) (90 thru 119 = 10) (120 thru 149 = 11) (150 thru 179 = 12) (180 thru 209 = 13) (210 thru 239 = 14) (240 thru 269 = 15) (270 thru Highest=16

Age at first emergency admission for stroke (in years) was recoded as:

(0 thru 39 =1) (40 thru 49 = 2) (50 thru 59 = 3) (60 thru 69 = 4) (70 thru 79 = 5) (80 thru 89 = 6) (90 thru 99 = 7) (100 thru Highest=8)

The number of elective admissions prior to stroke (count) was recoded as

(1=0) (2=1) (3=2) (4=3) (5=4) (6=5) (7=6) (8=7) (9=8) (10=9) (11 thru Highest=10)

The number of emergency admissions prior to the stroke (count) was recoded as

(1=0) (2=1) (3=2) (4=3) (5=4) (6=5) (7=6) (8=7) (9=8) (10=9) (11 thru Highest=10)

The number of emergency admissions after the stroke emergency admissions was recoded to

(1=0) (2=1) (3=2) (4=3) (5=4) (6=5) (7=6) (8=7) (9=8) (10=9) (11 thru Highest=10)

The number of emergency admissions after the stroke emergency admissions was recoded to

(1=0) (2=1) (3=2) (4=3) (5=4) (6=5) (7=6) (8=7) (9=8) (10=9) (11 thru Highest=10)

The Stroke diagnosis code of the FEAS was recoded to

'G45' = 'TIA'

'I60' and 'I62' = 'HAEM'

('I63','I65','I66') = 'OCCL'

'I64' = 'STRK'

The number of GP event dates prior to the FEAS was recoded as

0 =0

1 to 16 = 1

17 to 40=2

41 to 65 = 3

66 to 110= 4

111 to 179=5

>179 = 6

The number of statin prescription dates (count) and the number of anticoagulant prescribing dates prior to the stroke event were recoded as follows

0 to 1 = 0

2 to 28 = 1

29 to 182 = 2

183 to 364 = 3

>364 = 4

Blood pressure monitoring (count of events) was recoded as

0 = 0

1 to 4 = 1

5 to 8 = 2

9 to 12 = 3

13 to 16 = 4

>16 = 5

Recordings about smoking habit (count of events) was recoded as

(MISSING=0) (1 thru 5=1) (6 thru 10=2) (11 thru Highest=3)

Appendix 5: List of questions requiring work for continuation of this project

1 Inclusion of data from other datasets such as accident and emergency, outpatients, screening programmes and disease specific registers.

2 Use of the frequency of hypertension related primary care recordings over time

3 Development of prescribing indicators.. There are potentially many ways to refine any analysis of outcomes based on when the drugs were prescribed, which drugs were prescribed, and how that changed over time. Such refinement would require clinical input to provide guidance on the equivalence and appropriateness of the drugs. This is true of the lipid lowering drugs, anticoagulant drugs, and Nicotine Replacement Therapy drugs utilised in the analysis.

4 BMI and Weight measurements over time. Utilise a series of measurements over time for some patients and perhaps establish a group for whom weight is e.g. increasing or decreasing over time ie and indicator of success in weight management.

5 Finalisation of indicators from secondary care, overall length of stay vs average length of stay, number of readmissions, elective, emergency, should any ‘other admission’ types be considered?

6 Inclusion of changes to the secondary care reporting – there appears to be a difference pre and post 2008 in terms of ‘other’ methods of admission.

7 What indicators from primary care. These include Read diagnostic codes, but these do not match the ICD10 codes used in secondary care. Rules would need to be devised as to which codes are most reliable and should be used in case of diagnosis conflict between recording systems

8 How much does the reporting in to primary care from secondary care agree with the recording in secondary care? What is the mechanism for this recording – manual entry from discharge letters or electronic upload? Are the recording methods consistently in place across all GP practices?

9 How to decide appropriate prescribing to use as a treatment for each disease? This is a potentially large piece of work involving not only the recognition of drugs from a long list, but the changes in prescribing regimes over time for two reasons, the development of new drugs, and the changes in prescribing due to adverse reactions of the patients. Then the frequency of prescribing over time, and how this is considered needs to be established, where the drug is a long term preventative treatment.

10 The use of lifestyle factor information. So far it is not clear that we have captured all the potential Read codes that could be filled in relating to smoking and obesity, and methodologies need to be adapted from elsewhere to work out the meaning or contribution that multiple (varying) recording over time might make.

11 It has become clear that there is a difficulty in finding a suitable way to establish an indicator for ‘primary care activity’ from the data, due to the variety of sources feeding into the data input.

12 If a measure of ‘primary care activity is developed. It would be necessary to devise a way to work out a rate for those people only registered with a participating GP over a partial period in the Project timeframe

13 Does a change of GP in the Project time frame have an effect on the level of recording? This might be down to the transfer of GP records from one GP practice to another, and down to initial assessment interviewing by the GP seeing the patient for the first time

14 In a multi-dataset environment, some data may be irrelevant. How do we choose this? E.g Outpatient attendances are recorded by treatment function code. Should something relating to a totally different treatment function be considered on the patient pathway or should it be ignored, does this vary according to when it occurred? (Any interaction with medical professionals just prior to an emergency event could be relevant to changing a patient pathway.

Future projects might consider using a combination of the two clustering methods, using the hierarchical approach to identify the optimal number of clusters then the k-means approach to identify the most robust solution selecting that number of clusters. To establish the optimal solution, the variance in the indicators explained by the solution (a) the F-statistic and b) the Between Cluster Sum of Squares) could be plotted against the number of clusters. This is known as the ‘Elbow’ method.

15 For the purposes of this experimental use of cluster analysis, type of stroke was included as a variable to examine whether this contributed to the understanding of the clusters; future cluster analysis could be done for each type of stroke separately.

16 How does this piece of work fit in with other work going on elsewhere? There is considerable interest in modelling patient pathways as well as specific research into strokes. Further work should link in with other projects to check for duplication of effort and consistency of findings

17 The Chronic Conditions Management Framework was introduced in 2007 by WG, and there have been differences in management of diseases since this time so we already have changes in patient pathways going on.

18 Use of deprivation scores grouped by tenths in terms of severity of deprivation rather than equal tenths of population as has been utilised so far in the Project.

19 Migration and deprivation score: How does WIMD change with migration. This could be particularly important in the more elderly groups where retirement to another area, loss of partner to death, movement into supported accommodation etc may mean that there are a lot of people being classed differently because they moved to an area with a different WIMD score.

20 The significant variation between males and females in terms of age on admission and length of stay suggests that further work might complete cluster analysis for men and women separately. If different pathways exist for stroke by gender and some co-morbidities but not others, can a methodology be defined to establish this?