Dadansoddi ar gyfer Polisi

YMCHWIL GYMDEITHASOL Y LLYWODRAETH

GSR

GOVERNMENT SOCIAL RESEARCH

Analysis for Policy

Llywodraeth Cymru
Welsh Government

Ymchwil gymdeithasol

Social research

Number: 75/2013

www.cymru.gov.uk

Examining the Feasibility of Establishing a Wales Longitudinal Study

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

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

Rhys Davies WISERD
Cardiff University
46 Park Place
Cardiff
CF10 3BB

For further information please contact:

Name: Sarah Lowe

Department: Knowledge and Analytical Services

Welsh Government

Cathays Park

Cardiff

CF10 3NQ

Tel: 02920 826229

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

Welsh Government Social Research, 2013

ISBN 978-1-4734-0670-4

Table of Contents

Acknowledgements

I wish to record my thanks to colleagues from both the academic community and UK National Statistical Institutes who gave their time to be interviewed during the course of this project. I am particularly grateful to colleagues responsible for the development and support of the three existing UK Longitudinal Studies for their invaluable insight and hospitality. Within the Welsh Government, I am grateful to Sarah Lowe for her advice and guidance throughout the project. None of these organisations bear any responsibility for the opinions, interpretation or analysis undertaken in this report.

Chapter 1: Introduction

1.1 Aims and Objectives of the SimWales Project

The overall aim of the Welsh Government Programme to Maximise the Use of Existing Data (Data Maximisation) is to identify how ambitious the Welsh Government can be with plans to maximise the use of existing data and in particular with using linked data to provide enhanced data for the Welsh population. The aim of the 'SimWales' work stream is to examine the feasibility of establishing a Wales Longitudinal Study. As with the existing three UK Longitudinal Studies (the ONS England and Wales LS, Scottish LS and Northern Ireland LS), the inclusion of Census data is of considerable importance in terms of the usability of the WLS as a research resource. In common with the other Longitudinal Studies, the ideal outcome would be to develop a WLS which allows administrative records to be linked to Census Data. However, the proposals for a WLS extend beyond the establishment of a Census based longitudinal study in Wales that is similar in design to the existing studies. The proposals for a WLS differ in 2 key respects.

- Firstly, the ONS, Scottish and Northern Ireland Longitudinal Studies are based on sampling fractions of 1%, 5% and 28% respectively. The proposal for a WLS seeks to incorporate *100%* of Census records in order to support analysis of more detailed population sub-groups than has previously been possible from existing survey sources. Possible applications include more detailed levels of geographical disaggregation or the ability to distinguish groups that have protected characteristics under the Equality Act 2010. The 100% sample would, in theory, make it feasible to link consented survey data to be linked to the WLS in a way that has not been done for the existing UK based Longitudinal Studies.

- Secondly, in addition to linking administrative data sets to a census based template, SIM Wales aims to assess the feasibility of developing methods for the statistical matching of retrospective anonymised survey data to the census and administrative data template in order to create a simulated population of Wales. Statistical matching involves linking the survey records for an anonymous individual to the most similar individual within an administrative template on the basis of a variety of characteristics that are common to both data sources. For those members of the population who have not responded to a particular survey (or who have responded to a survey but who have not provided their consent to link), statistical matching could be used to create a simulated data set of the same magnitude as the administrative template (potentially, although not necessarily, complete coverage), but including the full richness of the survey data.

The aim of this report is to identify the key challenges that would need to be addressed to support the development of a Wales Longitudinal Study. The most fundamental issue relates to the role of the Office for National Statistics who, unlike in other devolved nations in the UK, have responsibility for conducting the Census in Wales and are custodians of Welsh Census data. This has practical implications in terms of where and how a WLS could be constructed. The desired sample size of 100% also has implications in terms of public acceptability and maintaining the confidentiality of respondents, although is necessary criteria if consented data from sample surveys is going to be integrated in to the WLS. Complete coverage of the population may therefore not prove to be feasible. An alternative approach to Census data could be to create a population level study based on an administrative template, such as registers of patients held by the NHS in Wales. However, the information held on administrative records is unlikely to be as rich as that collected by the Census. The aim of using statistical matching techniques within a WLS may therefore also be dependent upon the inclusion of Census data to generate a population spine that contains the required level of detail necessary to implement such techniques. The desired functionality of the WLS is therefore dependent upon achieving an integrated set of aims. The report examines the feasibility of achieving these aims and what options exist in terms of establishing a WLS.

1.2 Wider Context to the Research Programme

Developing Research Access to Administrative Data

The UK Strategy for Data Resources for Social and Economic Research (referred to as the National Data Strategy - NDS) sought to identify, prioritise and assist in the development of and access to research data. It provided a strategy for how different organisations could work with the research community to 'maximise the research potential of existing data and to create new resources, developing better access to existing data and facilitating a broad research agenda'. Supporting access to Census data is a key component of the NDS. The potentially disclosive nature of Census data (such as issues associated with small area data), the links that have been made to highly sensitive administrative records (such as those developed by the Longitudinal Studies) and the cooperation between organisations required in facilitating the construction of and access to Census based sources all underline the importance of supporting the infrastructure necessary to maximise the research potential of Census data.

The NDS also recognised the importance of administrative data for the purpose of research and to explore ways in which access to these resources could be enhanced. As a result, the Administrative Data Liaison Service was established to provide assistance for researchers who wish to make use of (or link to) administrative data held by different government departments. Towards the end of 2011, the Administrative Data Taskforce was formed by the Economic and Social Research Council (ESRC), the Medical Research Council (MRC) and Wellcome Trust to

'identify the factors inhibiting more widespread use of administrative data for research and to make recommendations for improvements'. The report[1] of the taskforce was published in December 2012 and included the following key recommendations:

1. A UK Administrative Data Research Network will be responsible for linking data between government departments. The proposed network will provide a single governance structure that will allow for consistent and robust decision-making.
2. An Administrative Data Research Centre (ADRC) should be established in each of the four countries in the UK.
3. Legislation should be enacted to facilitate research access to administrative data and to allow linkage between departments to take place more efficiently.
4. A single UK-wide researcher accreditation process, built on national and international best practice should be established.
5. A strategy for engaging with the public should be instituted.
6. Sufficient funds should be put in place to support improved research access to and linkage between administrative data.

Within both Wales and Scotland, strategy documents are being, or have been, developed in support of the greater utilisation of administrative data and data linking for the purposes of research. Within Wales, NISCHR is in the process of completing a report *Maximising the Use of Routine Data for Research in Wales* (due to be published in March 2013) which outlines its plans for to 'broaden investment in approaches to utilise routine data for research within the field of health and social care. Within Scotland, following the results of a consultation exercise[2], two strategic documents were published in November 2012 to support access to and analysis of linked data as part of the Scotland-wide Data Linkage Framework[3]. In the documents *A Strategy for Improving Data Access and Analysis* and *Guiding Principles for Data Linkage* the Scottish Government provides a 'long term vision' for improving both the governance arrangements and technical capacity to 'securely and efficiently link statistical data'. The proposed Data Linkage Framework builds upon the infrastructure established via the Scottish Health Informatics Programme (SHIP)[4], an NHS based facility that supports analysis of Electronic Patient Records. As well as a technical infrastructure and governance arrangements, the SHIP programme incorporates researcher training and public engagement activities. The Scottish Data Linkage Framework is composed of four blocks, including

[1] http://www.esrc.ac.uk/funding-and-guidance/collaboration/collaborative initiatives/Administrative-Data-Taskforce.aspx
[2] http://www.scotland.gov.uk/Publications/2012/03/3260
[3] http://www.scotland.gov.uk/Topics/Statistics/datalinkageframework
[4] http://www.scot-ship.ac.uk/

- the Guiding Principles document to provide guidance to data controllers with respect to data linking;
- a Privacy Advisory and Ethics Committee to provide guidance with respect to data linkage projects;
- an Information Gateway and Data Linkage Service to support researchers in the development of data linkage projects, create linked data sets and to provide safe settings for the purpose of analysis; and
- a Steering Group to oversee the strategic direction of the Linkage Centre and Advisory Committee.

Within Wales, the Secure Anonymised Information Linkage (SAIL) Project, based at the Health Information Research Unit (HIRU) at Swansea University[5], has already developed considerable expertise and an international reputation in developing mechanisms that support linking a range of administrative and clinical data sources. Funded by the Welsh Government National Institute for Social Care and Health Research (NISCHR)[6], it has made important progress in data anonymisation, acquisition, linkage, quality management and analysis. HIRU recently became one of the four UK E-Health Centres of Excellence[7] for research into linking electronic health data. The focus of SAIL's work has therefore largely been on linking administrative health data sets, although more recently other administrative sources have been introduced, such as education data from the Pupil Level Annual School Census. Under its Data Maximisation Programme, the Welsh Government have introduced questions on several of its surveys to ask respondents to provide their consent for their survey data to be linked to other sources of administrative data held about them, with a view to such data being incorporated within the SAIL system. SAIL provides a clear focal point for data linking activity in Wales.

ONS Beyond 2011

The idea for a Wales Longitudinal Study is also taking place during a period where ONS is examining options for the Census as part of its Beyond 2011[8] programme. The programme has arisen out of concerns expressed regarding both the costs and the continued relevance of a decennial Census in a rapidly changing economy. A range of alternative options are being investigated by ONS, including surveys and combinations of short form and long form Censuses.

[5] http://www.swansea.ac.uk/medicine/ils/healthinformationresearchunit/
[6] NISCHR is the Welsh Government body that develops, in consultation with partners, policy on research and development to reflect the health and social care priorities of the Welsh Government. http://www.wales.nhs.uk/sites3/home.cfm?orgid=952
[7] Details for this call for funding are available at http://www.mrc.ac.uk/Fundingopportunities/Calls/E-healthCentresCall/index.htm
[8] http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html

However, given the increasing availability of administrative data and the significant reductions in costs that the use of such data sets could provide, the Beyond 2011 Programme is primarily focussing upon examining the statistical properties of various administrative data sets and whether such sources would meet the requirements of users. The assessment of alternative options and a recommendation as to the best way forward is due to be made during 2014. The final decision rests with the UK Government and Parliament. At the time of writing, a public consultation has been launched which has invited users to think of the relative merits of two options for taking the census in the future[9]:

- A census once a decade, like that conducted in 2011, but primarily online.
- A census using existing government 'administrative' data supplemented by a 4% rolling annual survey.

In published Options Reports for the Beyond 2011, the ability to continue the existing ONS LS is specified as one of the criteria against which the ability of these options to meet the requirements of users is being assessed. Only the full census option allows for the existing LS to continue in its present form (i.e. the inclusion of socio-economic data from the Census), and possibly be enhanced through the incorporation of additional administrative data sets in the future. The second option does not allow for the inclusion of additional 'census-type' socio-economic data in to the LS beyond that extracted from the 2011 Census. This second option only allows for the existing LS to be further enhanced by the continuing addition of the life events data that already contribute the LS and the possible inclusion of additional administrative sources[10].

From a Wales perspective, it is important to understand the UK context to the Beyond 2011 programme. As will be discussed further in this report, the ONS is responsible for conducting the Census in both England and Wales. In both Scotland and Northern Ireland responsibility for the Census is devolved and the statistical offices in these devolved administrations will be undertaking their own reviews of options for the Census moving forward, subject to harmonisation requirements established in an Inter Administration Agreement between the three UK Census Offices[11]. Whilst the Welsh Government and other users of the Census in Wales have been consulted to ensure that future development of the England and Wales Census meets their requirements, the Beyond 2011 Programme is essentially investigating alternative approaches to the Census that would be implemented in both England and Wales. The Beyond 2011 consultation document also reaffirms

[9] http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-consultation/index.html
[10] http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-options-report-2--o2-.pdf
[11] Available from http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/background-to-beyond-2011/beyond-2011---the-uk-context/index.html

ONS's commitment to confidentiality. *Maintaining confidentiality is…….a fundamental principle of the Beyond 2011 Programme. Although administrative data may be used extensively in future, any data held will be securely stored and tightly controlled.*[12] Whilst the Beyond 2011 programme clearly places greater emphasis on data linking and ensuring that any legal arrangements are in place to allow for the sharing of data that is necessary to construct the new Census, the Beyond 2011 programme will not directly support the establishment of a Wales Longitudinal Study.

It is also important to note that the Beyond 2011 Programme will have an influence upon the content and structure of the existing Longitudinal Studies following the next Census in 2021. Nonetheless, the 2011 Censuses, whose inclusion within the three existing UK Longitudinal Studies is being validated at the time of writing, will remain the bedrock of socio-economic data that researchers using these studies will draw upon for the next decade. The decision to construct a Census based Longitudinal Study initially involves looking backwards rather than forwards. The Welsh Government holds an accumulated stock of administrative data whose value can potentially be enhanced by enabling links to be made with Census data that pre-dates it. For example, is it worthwhile to allocate parental social class as measured in the 2001 Census to National Pupil Database data held for Welsh children from 2004/5 onwards so that educational outcomes that were observed among Welsh children during the last decade can be better understood? In deciding this, it would have to be acknowledged that any WLS that could be created on the basis of existing Census data would probably change its form following the 2021 Census. Chapter 2 of this report touches upon how the Beyond 2011 Programme may impact on the development of the three UK Longitudinal Studies.

Devolution of Political Powers

The establishment of Census based Longitudinal Studies by ONS, NRS and NISRA each relies on the fact that these statistical offices have responsibility for conducting and administering the Census. Whilst ONS is responsible for undertaking the Census in England and Wales, NRS and NISRA are the custodians of their Census data and have therefore been in a position to develop longitudinal studies in a way that has not been possible in Wales. The main area of research that is conducted using these Longitudinal Studies relates to public health, an area under which law making powers have been devolved to the Welsh Government under Section 5 of the Government of Wales Act 2006[13]. The inability to develop a longitudinal database that is capable of contributing to the evidence base in Wales to inform the decisions of policy makers in a way that has been

[12] See page 6 of http://www.ons.gov.uk/ons/about-ons/user-engagement/consultations-and-surveys/archived-consultations/2012/beyond-2011---public-consultation/index.html

[13] http://www.assemblywales.org/bus-home/bus-legislation/bus-legislation-guidance/bus-legislation-guidance-documents/legislation_fields/schedule-5.htm

achieved in Northern Ireland and Scotland would appear to be inequitable. The particular circumstances faced by the population of Wales in terms of levels of real unemployment[14] and poverty[15] call into question the appropriateness of extrapolating evidence derived from national studies in order to inform decision making in Wales.

1.3 Structure of the Report

The remainder of this report is structured as follows. Chapter 2 provides an overview of the three existing UK Longitudinal Studies. The chapter examines why and how these studies were established; provides an overview of their structure, content and functionality; and describes how the researchers are supported in gaining access to these databases. Chapter 3 explores issues surrounding the inclusion of Census data within a Wales Longitudinal Study. With ONS having responsibility for the Census in Wales, the co-operation of ONS is clearly important if Wales is to establish a Census based Longitudinal Study and how this would be likely impact upon how and where a WLS could be constructed. The use of Census data as a population spine provides one option regarding the construction of a WLS. However, an alternative method could be to use administrative health records as a template against which other sources of data are linked, as is the case with the Northern Ireland Longitudinal Study. Chapter 4 therefore reviews recent research that has been undertaken to examine the completeness of population data derived from administrative health records. Secondly, the chapter describes the potential importance of the Annual Population Survey as a source of longitudinal socio-economic data for Wales. Chapter 5 discusses practical issues surrounding the implementation of statistical matching techniques within the WLS. These issues and the performance of statistical matching techniques are explored via an illustrative analysis which matches information on poverty from the Households Below Average Income (HBAI) data set on to the Labour Force Survey (LFS). Chapter 6 summarises the main findings of the project and makes specific recommendations.

[14] Beatty, C., Fothergill, S. and Gore, T. (2012) ***The real level of unemployment 2012***. Sheffield: CRESR, Sheffield Hallam University.
[15] Parekh A. and Kenway P. (2011). *Monitoring Poverty and Social Exclusion in Wales*. Joseph Rowntree Foundation.

Chapter 2: Overview of Existing Longitudinal Studies in the UK

2.1 Introduction

The SimWales fellowship seeks to establish the mechanisms that would be required to construct a 100% Sample Welsh Longitudinal Study that utilises the Census as the template or 'spine' against which other data sources are either linked or matched. The construction of such a database will face considerable challenges in terms of the agreement and development of mechanisms that would allow for it (who does the linking, where the linking is done and how it is done). Beyond the construction of the database, mechanisms will also need to be established regarding how users can access the data and how this access will be supported. A key partner in this respect is the Office for National Statistics (ONS). Unlike other devolved administrations in the UK, the Welsh Government does not have responsibility for conducting the Census in Wales, with ONS being responsible for conducting the England and Wales Census.

Although there are significant barriers that need to be overcome, the development of the WLS can benefit from drawing upon the experiences of three Longitudinal Studies that have already been established in the UK; the England and Wales Longitudinal Study (LS), the Scottish Longitudinal Study (SLS) and the Northern Ireland Longitudinal Study (NILS). In each case, links are made between Census data and primarily administrative health data. The circumstances and mechanisms through which these databases were established varied, as did the provisions that were put in place for supporting their use of these databases. Nonetheless, in each case considerable attention was given to establishing the aims and objectives of these databases and associated governance arrangements. The experiences of both data custodians based at the respective statistical offices and those based at the Data Support Units who have responsibility for supporting users of these databases each provide a useful insight in to the issues that will need to be addressed in the establishment of a WLS.

This chapter provides an overview of the three Census based Longitudinal Studies that have already been established in the UK. In addition to 'desk-based' research, interviews were conducted with staff from both statistical officers and DSUs. Section 2.2 provides an overview of the ESRCs Census Programme. This funding scheme has largely been the mechanism through which academic access to the three Longitudinal Studies has been supported. Sections 2.3, 2.4 and 2.5 respectively consider the Scottish, Northern Ireland and England and Wales Longitudinal Studies. Section 2.6 reviews available evidence regarding the costs and benefits of the Longitudinal Studies. Section 2.7 provides concluding comments.

2.2 ESRC Census Programmes and the UK Longitudinal Studies

The main overarching objective of the ESRC Census Programme is to provide academic access to UK census data through a framework for data acquisition, user registration, access software and expert support. Four data units were supported during the 1991 Census Programme, including the Longitudinal Study Support Unit (LSSU) at City University which supported access to the ONS Longitudinal Study, albeit via a safe setting based at the ONS[16]. In contrast to previous phases of the programme, the 2001-5 Census Programme was almost entirely service orientated since it only really supported the financing of what were referred to as Data Support Units rather than Research Units. This change in emphasis was the result of the decision to fund methodological innovations in other ESRC investments such as the Research Methods Programme, as well as through response-mode submissions via the Research Grant Board. Six DSUs were funded through this part of the Programme. Under this Programme, support for the ONS Longitudinal Study transferred to the Centre for Longitudinal Study Information and User Support (CeLSIUS) at the London School of Hygiene and Tropical Medicine. The structure of the Census Programme in the 2006-2011 funding phase was relatively similar to that seen between 2001 and 2005, with the same DSUs being supported following a competitive bidding process. In addition to this, two further DSUs were added to the Programme after 2006. Firstly, Northern Ireland Longitudinal Study Research Support Unit (NILS-RSU) was incorporated into the Programme in April 2009. Funding for a user support service for the Scottish Longitudinal Study (developed by the Longitudinal Studies Centre – Scotland) was also provided by the Programme from July 2009. The evolution of these support units is discussed in the remainder of this chapter.

In 2012, a new UK Data Service was funded by the ESRC for a period of 5 years[17]. The service was created largely via the integration of the Economic and Social Data Service, the Census Programme and the Secure Data Service. However, the support services provided for the three longitudinal studies were not included within the new UK Data Service. Instead, the ESRC established Research Support Units for the England and Wales (University College London), Scotland (University of St Andrews) and Northern Ireland (Queen's University Belfast) Census Longitudinal Studies to provide support and promote the studies within the academic, policy, and practitioner communities. To support these units, a UK Census Longitudinal Study Development Hub, based at the University of St Andrews) was also established to support and improve harmonisation across the three Census Longitudinal Studies.

2.3 England and Wales Longitudinal Study, Office for National Statistics

The ONS LS is the longest running of the UK Longitudinal Studies, being established in the 1970s to provide better data on occupational mortality and fertility. Only limited socio-economic data is

[16] The first user support service for the LS was actually set up in 1985 via a specific ESRC grant to John Fox
[17] http://www.esrc.ac.uk/news-and-events/press-releases/22231/new-national-digital-repository-for-social-and-economic-data.aspx

collected via the registration of births and death and it was felt that more reliable and consistent data could be created by merging data on life events with information related to the households of these same individuals extracted from Census records. The original sample was drawn in 1974 based upon 1971 Census data (England and Wales) to include every person in the census who was born on 1 of 4 specific days in the year and covers about 1% of the total population. The choice of a 1% sampling fraction was not arbitrary, but was based upon achieving a sufficiently large enough sample of individuals (estimated to be approximately half a million people) upon which estimates of occupational mortality could be based. Census information is also included for all people living in the same household as the LS member, although the LS does not follow up household members in the same way from census to census.

The same four dates have been used to update the sample during subsequent censuses The ONS Longitudinal Study is therefore a complete set of census records for individuals, linked between successive censuses and includes information on over a million study members (those born on the LS sample dates). The work that underpins linkage of data from life events and NHS registration events is carried out on behalf of ONS by the Health & Social Care Information Centre (HSCIC). The HSCIC maintains records of all registered users of NHS services and receives notifications of births, deaths and cancer registrations. Details of LS members are sent to the HSCIC to be identified and flagged as an LS member. Events occurring to LS members can therefore be identified and updated on the LS. Responsibility for the team that carry out this work moved from ONS to the HSCIC in April 2008, resulting in the transfer of staff between the two organisations.

User support for the ONS Longitudinal Study is currently provided by CeLSIUS, who are based at University College London[18]. The central task of CeLSIUS is to enable researchers to analyse LS data which is constructed and hosted at the Office for National Statistics (primarily via the Virtual Micro-data Laboratory based at Drummond Gate). Prior to 2012, user support for non-academic users was provided separately by ONS staff. However, since 2012 CeLSIUS is now also responsible for providing user support to the UK statutory and voluntary sectors as well as UK academics, although it is envisaged that some existing non-academic users will continue to be supported by ONS where it is sensible to do so. Responsibility for the construction of the LS lies with the Longitudinal Study Development Team (LSDT) within ONS Originally based in Drummond Gate with responsibility for running the dedicated LS safe-setting, this team are now located at the ONS office in Titchfield. This team remains responsible for supporting ONS based users.

Whilst the inclusion of additional life events data continues to maintain levels of interest in the LS, the opportunities for analysis provided by the addition to 2011 Census data to the LS clearly represents the next step shift in interest that will surround this data source. At the time of writing

[18] http://www.ucl.ac.uk/celsius

ONS, have completed their integration of 2011 Census Data in to the Longitudinal Study and have completed their own 'Alpha Tests' of the newly updated LS database. Towards the end of the 2012, ONS invited applications from LS users to undertake analysis of the database for the purpose of Beta Testing. These applications were reviewed in December 2012 and the ten successful projects were notified in January 2013. The Beta Test period will run until the autumn of 2013[19]. From a Welsh perspective, it is of interest to note that one of the essential criteria to be met by the portfolio of research projects selected for the purposes of Beta Testing was that at least some projects would encompass Wales-level analysis. The completion of these research projects should therefore provide a clearer indication of the potential of the ONS LS with respect to Welsh level analysis[20]. Access to the wider research community should take place following a launch event planned for November 2013 where the results of the Beta Testing projects are presented.

In comparison to the Scottish and Northern Ireland Longitudinal Studies (and other longitudinal and cohort studies in the UK), opportunities for making innovative links to other sources of data using the LS appear to have been more limited[21]. For example, the Scottish Longitudinal Study, the English Longitudinal Study of Ageing and the Millennium Cohort Study have each been linked to Hospital Episodes Statistics, a data warehouse containing details of all admissions, outpatient appointments and A&E attendances at NHS hospitals in England[22]. The UCL CeLSIUS team have identified linking to HES as a potential future line of enquiry regarding the further development of the ONS LS and would seem to represent a natural next step in the evolution of the ONS LS if issues surrounding data sharing can be resolved. ONS has made efforts to link additional administrative data to the LS in the past, notably data from DWP regarding claims for benefits relating to unemployment (LS project number 40011[23]). Originally, it was established that there was no legal gateway that allowed these DWP data to be used in this way. Renewed attempts were made after the Statistics & Registration Service Act 2007 came into effect. However ONS have had limited opportunities to establish new data sharing arrangements using the new legislation, and other priorities have taken precedence. Such attempts to link data to the ONS LS suggest that it is possibly easier to link data within the devolved administrations where data custodians are often co-located in the same government department.

Although research use of the ONS LS is relatively concentrated among academics based within London, CeLSIUS does support a wider geographical spread of the users compared to both the NILS and the SLS where users are understandably concentrated among a limited number of institutions. The development of aggregated data sets for customers in order to support off-site

[19] http://www.ons.gov.uk/ons/about-ons/who-we-are/services/longitudinal-study/2011-census-link-beta-test/index.html
[20] See section 3.6 for a description of selected Wales specific LS projects.
[21] See Calderwood, 2006, Figure 2. https://www.iser.essex.ac.uk/files/survey/ulsc/methodological-research/mols-2006/scientific-social-programme/papers/Calderwood.pdf
[22] http://www.hscic.gov.uk/hes
[23] http://celsius.lshtm.ac.uk/projects/plattEthnicity.html

analysis, remote job submission and the development of downloadable tables for potential users to consider the feasibility of analysis have all been important areas of CeLSIUS's work. The UCL team are exploring the development of non-disclosive synthetic data sets so that researchers can develop statistical programmes without having to visit the safe setting.

2.4 Scottish Longitudinal Study – National Records of Scotland and University of St Andrews

Although originally included in the plans for the establishment of the ONS Longitudinal Study, Scotland was subsequently dropped from the study when it became apparent to the separately funded General Register Office for Scotland (GROS) that a 1% sample would not be sufficient to support the analysis of Scottish-specific studies. The absence of a Scottish Longitudinal Study however became a cause for concern over a number of years in the context of higher morbidity and mortality rates in Scotland compared to England and Wales ('the Scottish Effect'). This prompted a group of academics to seek funds from the Scottish Higher Education Funding Council to re-establish a Scottish Longitudinal Study, which was completed in 2004. Unlike the ONS LS, the academic community largely provided the impetus to construct the SLS database[24]. With funding from a variety of sources, the Longitudinal Study Centre – Scotland (LSCS)[25] based at St Andrews University was established to create and maintain the Scottish Longitudinal Study (SLS).

In Scotland, the National Records of Scotland (NRS, previously GROS) has responsibility for conducting the Census[26]. In developing the SLS, it was decided that the data and access to the data should be based at the then offices of the GROS in order to improve perceptions in the face of public concerns regarding data security. The LSCS team have a split location. Half of the team are physically located at NRS offices to work on data development. The other half of the team work at St Andrews, although when they provide user support they travel to Edinburgh. All LSCS staff are employees of St Andrews University with the exception of one who is supported by the NRS. As ESRC funding has only previously covered the provision of support for academic users, the NRS staff member notionally supports government users whilst the St Andrews User Support Officers (USOs) support academics. In practice, there was a degree of cross over with all USOs providing support to academic and non-academic users from time to time. Under the new funding arrangement, USOs will support both academic and government users, which is regarded as helpful in terms of developing links. The NRS employee is the data custodian and the point of

[24] See Issue 1 of CeLSIUS newsletter for short description of the establishment of the SLS.: http://celsius.lshtm.ac.uk/documents/newsletter001.pdf
[25] http://www.lscs.ac.uk/
[26] The responsibility of Registrar General for Scotland for conducting the census in Scotland was established under the 1854 Registration of Births, Deaths and Marriages (Scotland) Act.

liaison with the LSCS team. This is perceived as important for the NRS as they have an employee at the heart of SLS operations.

The SLS currently holds census data for 1991 and 2001 based upon a 5.3% sample of the Scottish population (those born on one of 20 birth dates). This yielded a sample size of approximately 275 thousand people who were extracted from the 1991 Census. Linked Census data for 2011 is due to be incorporated within the SLS by the autumn of 2013. Defence against accidental disclosure was not regarded as a key determinant of the size of the existing SLS sample. Although smaller than the 500 thousand than that associated with the ONS LS, this is regarded as 'perfectly adequate' for a majority of research projects with an acceptance that there would inevitably be projects related to particularly small population sub-groups or certain rare events that could not be supported by the SLS. Given resources, LSCS staff suggest that a larger SLS would have to utilise more probabilistic matching techniques resulting in lower levels of accuracy.

Opportunities for developing new links to the SLS appear to be greater than those that exist for the ONS LS. SLS links to hospital episode statistics, enabling links to be made to information about mental and physical illnesses as well as information regarding maternity. However, for health data only extracts necessary for specific projects are pulled in to the SLS on a project by project basis. The data have also been linked with information on marriages (Scotland uniquely collects date of birth information on marriage certificates) and from the Annual Schools Census. The digitisation of the 1939 National Register and the indexation of vital events data back to 1855 by the NRS also provide further opportunities for data linking27. The SLS has not been linked to surveys conducted by the Scottish Government due to the small samples that would emerge.

The SLS is relatively risk averse in its operation. In contrast to the ONS LS, the LSCS do not construct and release aggregated data sets for off-site analysis. Users tend to come in to the safe setting initially to gain familiarity with the data. Statistical programmes are then developed off-site which are then submitted to USOs thereafter. Whilst there are several examples of projects conducted across different HEIs in Scotland, academic use of the SLS is relatively clustered in St Andrews and Glasgow Universities due to the concentration of expertise and the ability to access the data. Now that further funding has been secured, the LSCS are looking to explore the possibilities of making greater use of aggregated data sets to widen access to the SLS. However, NRS is cautious with regards to this approach, indicating the significant development work would be required to think about how access to such data would be provided. The costs and benefits of such an approach would also have to be balanced against alternative approaches to widening access. The SLS team suggested that a more likely route to expanding access in the short term is to produce synthetic data sets rather than trying to expand access to the real data. Such data sets

[27] The details of an ESRC funded feasibility study which examined these links is available at: http://www.esrc.ac.uk/my-esrc/grants/RES-348-25-0014/read

could be given to researchers through the Secure Data Service for the purposes of supporting remote job submission. Expansion through the ONS VML was regarded as a potential medium term solution, particularly with respect to enabling researchers to access the 3 UK longitudinal studies from the same location. The use of an experienced and trusted safe setting such as the VML would allow the LSCS to assess demand to inform the direction of future developments. However, many practical questions would have to be addressed, not least whether the data remains on an NRS server or gets transferred to ONS.

In terms of Beyond 2011, it was felt that the potential movement towards an administrative based Census represented the beginning of a period of change for the SLS, and would possibly determine what the SLS would like in the future. Although the format is unclear, the LSCS and NRS are well placed in terms of having the skills, experience and governance arrangements in place that are necessary to work with linked administrative data. Whilst the availability of new administrative sources could provide new opportunities, it is felt that the possible absence of data on household/family structure, housing conditions and occupational coding would be an issue.

2.5 Northern Ireland Longitudinal Study, Northern Ireland Statistics and Research Agency

The Northern Ireland Statistics and Research Agency (NISRA), an Executive Agency within the Department of Finance and Personnel (DFP), is the official statistics and research agency for NI Government. Among the aims of NISRA are to provide statistical research and information into Northern Ireland, analytical support in the development of policy and registration services to the public. NISRA also has responsibility for the Census of Population in Northern Ireland, although in practice NISRA join up with ONS in terms of questionnaire development and data capture[28]. As within Scotland, the impetus for a Northern Ireland Longitudinal Study primarily came from the academic community among which a number of people lobbied NISRA to start a NI Longitudinal Study. NISRA recognised the benefits that could be gained from linking data from different sources and in particular the importance of longitudinal data. In the context of the existing ONS and Scottish Longitudinal Studies and the relative absence of such data in Northern Ireland, in 2003 NISRA developed a business case to support the establishment of a Northern Ireland Longitudinal Study by joining information from various administrative databases across Northern Ireland.

[28] For an overview of the history of the Census in Ireland and Northern Ireland, see: http://www.nisra.gov.uk/archive/demography/publications/annual_reports/2011/Chapter2.pdf

The case for support primarily considered 2 options for the NILS; one based on the 2001 Census (referred to as the 'Pilot LS') and a second that incorporated both the 1991 and 2001 Census (referred to as the 'Full LS'). The inclusion of the 1991 Census data would have been a resource intensive task as the 1991 Census assured respondents that names and addresses would not be held electronically and therefore the original paper forms would have to have been interrogated. At the time the business case was developed, there was some uncertainty surrounding the resources that would have been required to undertake such an exercise. The Full LS option was therefore not considered in detail, but would remain an option to be returned to in the future if the Pilot LS could be successfully implemented. It was proposed that the Pilot NILS would contain a linked database which contained a sample consisting of a third of records from the 2001 Census linked to 1) vital event registrations from the General Register Office; 2) Health and Personal Social Services (HPSS data) and 3) cancer registrations from the NI Cancer Registry dataset.

The recommendations of the business cases were accepted. However, additional legislative changes were required in order to establish the NILS. Census legislation for Northern Ireland is based on the Census Act (1969) Northern Ireland, which was broadly similar to the Census Act (1920) GB. However, legal advice from the Office for National Statistics indicated that their ability to conduct the Longitudinal Study in England and Wales was based on a part of the GB legislation that was not replicated in the NI Act[29] and that an amendment to section 5 of the Census Act (1969) Northern Ireland would be required if NISRA wished to undertake a study in Northern Ireland. Therefore an additional section to Law Reform Miscellaneous Provisions Order (2004) was required to amend section 5 of the Census Act (1969) Northern Ireland to bring it into line with the Census Act (1920) GB.

NISRA was entirely responsible for the construction of NILS. The continuing creation, maintenance and development of the NILS are the responsibility of the NILS-Core team. The NILS-Core is part of NISRA's Demography and Methodology Branch who report to the Head of Demography. Initially, NISRA was also responsible for providing a research support function. However, increasing demands on access to NILS put extra pressure on NISRA staff. NISRA therefore requested support from the ESRC for the establishment of a unit dedicated to supporting the use of NILS. The Northern Ireland Longitudinal Study Research Support Unit (NILS-RSU) was established in April 2009 with funding from the ESRC Census Programme. As is the case within

[29] Section 5 of the Census Action 1969 (Northern Ireland) stated that: *The Ministry of Finance may collect and publish from time to time any available statistical information concerning the population of Northern Ireland in the interval between one census and another.* Section 5 of the Census Act 1920 (GB) stated that: *It shall be the duty of the Registrar-General from time to time to collect and publish any available statistical information with respect to the number and condition of the population in the interval between one census and another, and otherwise to further the supply and provide for the better co-ordination of such information, and the Registrar-General may make arrangements with any Government Department or local authority for the purpose of acquiring any materials or information necessary for the purpose aforesaid.*

Scotland, NISRA have funded a member of NISRA staff to be based within the NILS-RSU so that the unit provided support to both academic and non-academic users. Core unit functions of the NILS-RSU are to promote the research potential of the data, to provide expert support services to researchers from the academic and government sectors who want to use the NILS data. Most outputs are derived from the direct analysis of micro-data by locally based researchers visiting the well used safe-setting at NISRA offices. NILS-RSU also provides a remote job submission service, although as is the case with Scotland, aggregated data sets are not released. NISRA are cautious about developments that would allow Census data to be released anywhere outside of NISRA and there are no plans for the LS to be accessed from locations other than the NILS safe setting.

The relative strength of the NILS is its sample size, with almost a third (the final NILS sample was 28%) of the population from the 2001 Census being included. NILS only includes 2001 Census data at present and so changes in key socio-economic measures cannot be measured, although 2011 data is currently being integrated and will be available for analysis (subject to the successful completion of Beta Testing projects) by the autumn of 2013. Discussions continue to take place with respect to the incorporation of 1991 Census data within NILS. Due to the legal gateways that provide access to the data (via health legislation that has a clause which states that health services data can be used for the purposes of health research), project applications to use the NILS must meet 2 criteria; a) a longitudinal element and b) a health element (although interpreted broadly via the WHO definition that also incorporates well-being). In terms of sample size, the main factor underpinning the size of the NILS was to achieve sufficient number of observations to support NI level analysis whilst still ensuring the anonymity of respondents and being sensitive to the perceptions of the public. It is also important to note that rules surrounding access ensure that the NILS cannot simply be used as a large SARS. The Northern Ireland Mortality Study (NIMS) does provide a 100% sample, but it only contains deaths data linked to the Census, with no other links being made to this database

The construction of the NILS has also enabled innovative links to be established with other sensitive heath data such as dental records, cancer screening and prescriptions. The provision of health and social care as an integrated service in Northern Ireland also enhance the availability of social care data sets. These projects are referred to as Distinct Linkage Projects and involve data, subject to the necessary ethical approval, being brought in to the NILS only for the duration of the project. NISRA staff emphasised the reluctance of data custodians to provide data for non-specific purposes and there was no desire to create a data warehouse or repository. However, once a data has been used for research, a precedent is set which should mean that it is relatively straightforward to gain agreement for that data to be used again.

NILS is regarded as being a well placed for Beyond 2011. If the current Census is replaced by an alternative based upon administrative sources, the skills, expertise, governance and administrative

arrangements that have been acquired during the development of NILS are regarding by the NILS-RU team as providing a reasonable 'step off' or 'model' for how the analysis of Census data will look in the future. In some respects, if Beyond 2011 provides the impetus of legislative changes that allow NISRA to acquire other data sets necessary for constructing population estimates for NI (e.g. schools, HE and benefits data), this raises the possibility that such data could be more readily incorporated in to the NILS. However, once again it was emphasised that research utilising any variables that were available from these sources that were beyond what would constitute the 'core' data items of an administrative based Census would probably be undertaken under the auspices of Distinct Linking Projects.

2.6 Costs and Benefits of Longitudinal Studies

There are a variety of benefits that have been associated with the development of the 3 UK Longitudinal Studies. Section 2 of the Scottish Government Document *A Scotland-wide Data Linkage Framework for Statistics and Research: Consultation Paper on the Aims and Guiding Principles*[30] has identified 6 benefits associated with data linking. Each of the benefits (listed below) is of potential relevance to the continuing activities of SAIL and the establishment of a Wales Longitudinal Study. It can be seen that the benefits listed are varied. Those most easy to quantify include the ability of data linkage to support the development of relatively low cost longitudinal research which is of particular importance in the context of being able to evaluate the impact of policy interventions. Evaluations of interventions, such as those covered by the New Deal Programmes, have often relied upon the collection of primary data which can be of considerable expense. Whilst administrative data cannot provide information about the perceptions of participants, the use of Counterfactual Impact Evaluation techniques (see Chapter 5) can provide a cheaper way of providing evidence as to the effectiveness of interventions. For example, in Northern Ireland the NILS data set has been used in the evaluation of how personal, socio-economic characteristics have influenced the take-up of screening services for breast cancer. The report highlights low levels of take-up are concentrated in deprived areas and how improvements in the organisation of the service could contribute to reducing socio-economic inequalities in screening.[31] Other benefits are more difficult to quantify. The Scottish Government place considerable emphasis upon the role of data linking in terms of enhancing the international reputation of research in Scotland and the effects that this can have in terms of attracting research income and employment creation.

[30] http://www.scotland.gov.uk/Topics/Statistics/datalinkageframework
[31] http://www.qub.ac.uk/research-centres/NILSResearchSupportUnit/FileStore/Filetoupload,288970,en.pdf

Benefits of Data Linking from *A Scotland-wide Data Linkage Framework for Statistics and Research: Consultation Paper on the Aims and Guiding Principles*

Benefit 1: Data linkage will help speed up cycles of improvement through the delivery of a higher quality cross-sectoral evidence base to inform public policy and strategic spending decisions

Benefit 2: Enable better use of existing data to develop efficient methods of producing demographic and census-type statistics

Benefit 3: Data linkage will increase the power of official statistics available to all

Benefit 4: Data linkage will allow relatively low cost longitudinal research to be conducted both retrospectively and prospectively, informing preventative spend

Benefit 5: Increase the capacity to robustly evaluate the costs, benefits and risks of new health, social, educational and associated programmes

Benefit 6: Data linkage will provide globally unique exemplars of research excellence, enhancing Scotland's reputation and attracting investment and job creation to Scotland

Whilst the benefits of data linkage are clear, they are however more likely to be realised where data linkage is supported by facilities that provide research access and some form of User Support. Many of the benefits listed above were also identified as being of importance in the context of the establishment of the Virtual Microdata Laboratory at ONS (e.g. the ability of government departments to improve the evidence base in a low cost way). However, the gains associated with ONS developing closer links with a community of researchers was also regarded of importance[32]. The provision of an access and user support facility enabled researchers to come into contact with those responsible for producing the sources, enhancing opportunities for knowledge transfer in both directions. Such benefits have also arisen from the establishment of the units that support access to the existing longitudinal studies. The development of communities of researchers increases the available pool of skilled labour that government can call upon to provide expertise and thereby enhancing the competitiveness of procurement exercises. Data

[32] Ritchie F (2008), Secure Access to Confidential Micro-data: Four Years of the Virtual Micro-data Laboratory: Economic and Labour Market Review, vol. 2, no.5. Available at: http://www.ons.gov.uk/ons/rel/elmr/economic-and-labour-market-review/no--5--may-2008/economic---labour-market-review.pdf

access and user support are important components in terms of realising the benefits associated with increased data linking activities in Wales.

It is difficult to estimate the costs associated with establishing and supporting a Census based Longitudinal Study. The business case for NILS estimated that he establishment of a longitudinal study based upon links made to a single Census would cost approximately £330k (at 2003 prices) during a 3 year period from 2003/4 to 2005/6. However, the ongoing costs associated with extracting, linking and supporting new sources of administrative data are more difficult to quantify. If the existing ONS LS was to be boosted for Wales, any new tracing and linking work that would have to be undertaken by the HSCIC in Southport would need to be funded. More up to date information on the costs associated with supporting users of these facilities can be found via the information held on the ESRC website about the grants awarded to the support units. Prior to the establishment of the new Research Support Units in 2012 (see Section 2.2), the existing support units were given 12 month extensions to their contracts during 2011/2012 in order to provide continuing user support whilst the new arrangements were being put in to place. The cost of these extensions were approximately £228k for CeLSIUS; £280k for LSCS and £110k for NILS. Variations in costs reflect the resources necessary to maintain existing levels of service provision and the varying levels of involvement of the statistical offices. However, it is noted that the development of a Census based WLS at ONS could benefit in part from the existing user support provided by CeLSIUS.

2.7 Concluding Comments

Given pressures to reduce expenditure on surveys, concerns regarding response rates to social surveys and the burden such surveys place on society, the continuing development of expertise in the construction and analysis of administrative data for research will be of key importance to researchers within the UK. The Longitudinal Studies and the research units that support their use are engaged in activities that have supported the development of skills and expertise that are of particular relevance in terms of the direction in which developments in data sources are moving, such as the Beyond 2011 programme. The ability to both establish and develop these databases has ultimately depended upon the long established devolution of responsibility for the Census and Registration Services. The Office for National Statistics is responsible for conducting the England and Wales Census. The co-operation of ONS would therefore be essential to the establishment of a Census based Welsh Longitudinal Study. Issues surrounding the potential involvement of ONS in a Welsh Longitudinal Study are considered in the next chapter.

This review of the operations of the 3 existing UK Longitudinal Studies provides a useful insight in to some of the issues that would have to be addressed in the establishment of a Census based Welsh Longitudinal Study. Several common themes emerge. Firstly, within both Scotland and

Northern Ireland, the impetus to establish Census based longitudinal studies came from their respective academic communities. The research interests of these academics were primarily related to public health, where the availability of longitudinal data is particular importance. Secondly, whilst the ability to link additional sources of administrative data is an important aspect of their functionality, those responsible for the data sets are careful to emphasise that they do not wish to create 'Big Brother' data warehouse where different sources of administrative data are accumulated. Linked data sets are developed for the duration of specific projects and are then dispersed upon the completion of these studies. Thirdly, there appears to be no appetite among the census offices for the sampling fractions of these databases to be increased. Within the context of concerns surrounding maintaining the anonymity of respondents, existing sample sizes are regarded as adequate for a majority of research questions requiring longitudinal data. These databases are not regarded as providing a means for improving data for small areas or minority groups.

Chapter 3: ONS and the Wales Longitudinal Study

3.1 Introduction

Compared to the other devolved nations of the UK, Wales does not have its own Census-based Longitudinal Study. Instead, the Welsh population is included in the ONS England and Wales Longitudinal Study (LS). The impetus for establishing the Longitudinal Studies was to facilitate analyses of health outcomes. An important characteristic shared by the ONS, Scottish and Northern Ireland Longitudinal Studies is the availability of a range of demographic and socio-economic variables which enables analysts to undertake more detailed analyses of factors correlated with health outcomes than that which would be possible based upon administrative data alone. Whilst the design, content, coverage and flexibility of these studies varies, it is the inclusion of Census data that provides the opportunity to explore health outcomes across a variety of measures and is crucial to their importance as a research resource.

In addition to their core elements of Census and vital events data, the Scottish and Northern Ireland studies have been successful in incorporating additional administrative information into their respective Longitudinal Studies. Within Scotland, hospital admissions data, education data for those in publically funded schools and marriage events can be linked into the Scottish Longitudinal Study. Within Northern Ireland, NILS has also established a process that allows for additional administrative data sets to be linked for specifically defined one-off studies through the Distinct Linkage Project (DLP) process. Examples of linkages undertaken so far include NILS and cancer screening, prescribing and dental services data. These are subject to additional legal and ethical scrutiny and privacy protection protocols. Within Wales, in terms of data linking SAIL provides similar functionality to that provided by the SLS and NILS. The key limitation within Wales is that Census data has not been incorporated in to SAIL. As discussed in Chapter 2, the ability of both Scotland and Northern Ireland to construct longitudinal studies that incorporate Census data relies on the fact that legal differences mean that responsibility for conducting and administering the Census has been devolved. Whilst ONS is responsible for undertaking the Census in England and Wales, NRS and NISRA are the custodians of their Census data and have therefore been in a position to develop longitudinal studies in a way that has not been possible in Wales.

This Chapter explores issues surrounding the Census that are relevant to the development of a Census based Wales Longitudinal Study. Section 3.2 outlines the legal framework under which the Census is conducted and the assurances given to respondents regarding confidentiality. Section 3.3 outlines the general approach of ONS to providing access to Social Survey micro-data. This sets the context for understanding the mechanisms used for providing access to Census micro-data, described in Section 3.4. Aside from the construction of its Longitudinal Study, the

ONS has frequently used Census for the purposes of data linking. Section 3.5 provides outlines the circumstances under which such linking activities are conducted. Two options are then explored with regards to enhancing Census based longitudinal data for Wales. Section 3.6 describes how use of the existing ONS Longitudinal Study could be enhanced. Section 3.7 explores some of the issues surrounding boosting the ONS LS for Wales. Section 3.8 concludes.

3.2 The Legal Framework of the England and Wales Census

The 1920 Census Act gave statutory authority to the Registrar General for England and Wales to conduct a census every five years, although there has traditionally been a ten year gap[33]. Given its compulsory nature and, by definition, its largely complete coverage of the population, maintaining the confidentiality of the England and Wales Census is of upmost importance to ONS.

> *Data security and confidentiality of personal information is a top priority for the Census. It is central to the design of robust systems, processes and legal arrangements with contractors. ONS confirms its overriding commitment to ensuring the confidentiality of personal Census data for a period of 100 years and its use strictly for statistical purposes only[34].*

A variety of statements published on the ONS website underline the commitment ONS to maintaining the confidentiality of Census respondents. These include sections dedicated to providing detailed information regarding how Census data is collected, processed and held. Legislative changes have meant that the wording of these commitments has changed between the 2001 and 2011 Censuses. The pages dedicated to guidance and methodology surrounding the 2001 Census underlines the provisions for security outlined in the 1920 Act:

> *The confidentiality of the information supplied by the public is protected by legislation. In Great Britain, the Census Act 1920, as amended by the Census (Confidentiality) Act 1991, and provisions set out in the Census Regulations lay down penalties for the unlawful disclosure of information from the census by anyone involved in taking a census. It is unlawful for the Census Offices to pass any census information to other Government departments or any other organisation except for the purposes of the Census Act itself or*

[33] See section 1(1)(c)(i): http://www.legislation.gov.uk/ukpga/Geo5/10-11/41

[34] http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/the-2011-census-project/commitment-to-confidentiality/index.html

the Public Records Act 1958. Under this latter legislation, the census returns are closed to public inspection for 100 years[35].

The introduction of the Statistics and Registration Service Act of 2007 (SRSA) transferred the Census and other statistical functions of the Registrar General for England and Wales to the Statistics Board (UK Statistics Authority) formed on 1 April 2008. As a result, the confidentiality provisions of the Census Act have now been replaced by the provisions contained in the SRSA. The role of this new legislation in maintaining the confidentiality of census respondents is summarized by ONS as follows[36]:

Census data confidentiality is protected by the Statistics and Registration Service Act 2007 (SRSA).

The confidentiality provisions in SRSA and the duty on the Board to maintain confidentiality in the Census in England and Wales have replaced the confidentiality provisions of the Census Act.

Section 39 of the SRSA prohibits the disclosure of personal information with a penalty of imprisonment for a maximum of two years, a fine, or both.

Respondents to all ONS surveys are provided with a number of assurances regarding how they will be treated and how their data will be used. These assurances are outlined in the Respondent Charter for Surveys of Households and Individuals which supplements all commitments made on *letters, leaflets, survey questionnaires and the ONS website, as well as those made on the phone or in person by our interviewers and other employees[37].* The charter includes a commitment to keep the information of respondents secure and confidential. The principals of the charter are directed by the Code of Practice for Official Statistics and the provisions contained in the Data Protection Act and the SRSA.

In legal terms, there is nothing inherently special about Census data compared to the many other surveys conducted by ONS. In practice however, given the compulsory nature of the Census and the aim of achieving a maximum response from the entire population of England and Wales, ONS is far more explicit regarding the security measures for the Census than those included in the Respondent Charter or the specific assurances given to respondents of other surveys. These

[35]http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/census-legislation/security-and-confidentiality/index.html
[36]http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/the-2011-census-project/census-data-confidentiality-and-uk-law/index.html
[37]http://www.ons.gov.uk/ons/guide-method/surveys/respondents/household/charter-for-household-survey-respondents/index.html

attitudes are reflected in more restrictive access arrangements surrounding census micro-data compared to micro-data from other sources.

3.3 Access Arrangements for ONS Social Survey Micro-Data

Whilst the ONS primarily produces aggregate National Statistics derived from surveys or administrative sources, the importance of providing access to the underlying micro-data that underpins these aggregate statistics is well established. ONS provides access to such 'unit record' information via a range of services so that researchers and analysts have the necessary flexibility to interrogate the data to undertake more detailed analyses than that which can be provided by facilities that provide aggregate statistics, such as Neighbourhood Statistics and NOMIS websites. Therefore, in accordance with Principle 1 of the UK Statistics Authority Code of Practice (CoP), micro-data files of different degrees of disclosiveness are made available to other government departments, academics and other researchers. Access to microdata and disclosive data, that is, data which have the potential to identify an individual record, requires the approval of the ONS Microdata Release Panel (MRP) before the data can be provided. The MRP ensures that the release of micro-data conforms to the disclosure provisions of underlying legislation, assurances given to respondents regarding their confidentiality and principles established within the code of practice[38]. The MRP also provides guidance for data depositors, published in *GSS Disclosure Control Policy for Microdata Produced from Social Surveys.*

The least disclosive, fully anonymised 'standard level' micro-data can be lodged at the UK Data Archive (UKDA) at the University of Essex under an End-User Licence (EUL) arrangement. Under the EUL, researchers can log on to the UKDA and are free to download the data to their institutional PC. Other than researchers being required to register their use of this data and providing a declaration that they have understood the terms and conditions under which access to the data is being provided, no formal application is required to gain access to EUL data. Use of the data is not time limited and no additional security arrangements are required to be put in place. More disclosive versions of social survey data sets, referred to as Special License data sets, are also lodged at the UKDA. The most common distinction between SL and EUL data is that the SL data sets typically contain more detailed geographical identifiers compared to EUL data[39]. However, SL data sets may also contain additional variables that are not included in all on EUL data sets or retain a more detailed level of disaggregation for other variables such as occupation, industry or ethnicity.

[38] GSS Disclosure Control Policy for Microdata Produced from Social Surveys, available at: http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-social-survey-microdata/index.html

[39] EUL data sets typically contain geographical information coded at the level of Government Office Region or similar.

The UKDA restricts access to SL data sets to users who have received the necessary authorisation from ONS via applications made under the Approved Researcher process, the mechanism provided by the SRSA for ONS to provide access to disclosive data[40]. Under the Approved Research process, all applications to access SL Data are scrutinised by the MRP. The application process requires researchers to demonstrate evidence of both their own competence and experience of handling sensitive data and to outline the statistical purpose for which they require the data. Applicants are asked to demonstrate why non-disclosive data from other sources (including EUL data) was not sufficient to meet their analytical requirements. The applicant also has to provide assurances regarding the physical and technical security arrangements under which the data will be held. Subject to the approval of the MRP, the UKDA provides the data to the applicant via an encrypted download. Applicants are also asked to specify the length of time for which they will require the data, after which they can either apply for an extension or inform the UKDA that the data has been securely destroyed.

Finally, two further mechanisms are available for providing secure access to data that are too detailed, sensitive or confidential to be provided under the EUL and SL licenses via the UK Data Archive. During the early part of the last decade, ONS established the Virtual Microdata Laboratory. The facility was established as part of the Business Data Linking project and it initially provided access to micro-data collected from ONS Business Surveys conducted under the 1947 Statistics of Trade Act. An inherent difficulty in providing access to business data is that such data is inherently disclosive, with large organisations being relatively straightforward to identify. To overcome these difficulties, the VML facility provides secure on-site access to disclosive confidential data to external researchers. Subject to applications being approved by the MRP and following the completion of a training course in the operation of the VML, researchers visit a technically and physically secure, safe and supervised 'laboratory' environment within ONS offices. All statistical outputs produced by the researchers are checked by the VML team for the purposes of disclosure control. Sanctions were put in place to penalise researchers who were found to be in breach of the conditions imposed by ONS (see Ritchie, 2008)[41.]

The nature of the data sets held in the VML gradually widened over time, to include both new business data sets, prices data, Census micro-data (discussed below), 'client file' versions of social survey data sets (more disclosive than SL data) and the ONS Longitudinal Study. Following the success of the VML, the ESRC established the Secure Data Service at the UKDA in 2011[42]. The operation of the SDS is broadly based on the procedures established and developed by the VML.

[40] A description of the Approved Researcher process is available at: http://www.ons.gov.uk/ons/about-ons/who-we-are/services/unpublished-data/access-to-ons-data-service/index.html

[41] Ritchie F (2008), Secure Access to Confidential Micro-data: Four Years of the Virtual Micro-data Laboratory: Economic and Labour Market Review, vol. 2, no.5. Available at: http://www.ons.gov.uk/ons/rel/elmr/economic-and-labour-market-review/no--5--may-2008/economic---labour-market-review.pdf

[42] http://securedata.data-archive.ac.uk/

The key difference is that researchers analyse the data remotely from their institutional desktop or in a safe room. However, data cannot be downloaded from the SDS and all outputs are checked by the SDS team for the purposes of disclosure control. Many of the data sets (and users) of the VML have now been 'migrated' across to the SDS. The VML, however, continues to operate for the purposes of providing a facility for government researchers who are not eligible to use the SDS and for specific data linking projects that require the processing of identifying data by ONS (e.g. incorporating externally collected information about businesses on to ONS business data sets).

3.4 Arrangements for Research Access to Census Micro-Data

Census Microdata

The increased emphasis placed upon maintaining the confidentiality of Census respondents has contributed to more restrictive and limited access to Census micro-data compared to other ONS surveys. Funded under the ESRC's Census Programme, the Cathie Marsh Centre for Census and Survey Research at Manchester University has provided access to a range of Census microdata from 1991 and 2001, collectively referred to as the Samples of Anonymised Records (SARS). The following products based on the 2001 Census are available.

- The Individual Licensed SAR (IL-SAR) – 3% sample containing data on a full range of census topics plus summary household level data. Geographical information is given down to Government Office Region.
- The Special License Household SAR (SL-HSAR) – 1% sample containing data on over 200,000 households and 500,000 individuals living within those households. No geographical identifiers are available. Available via the UKDA.
- The Small Area Micro-data (SAM) file - 5% sample containing more detailed geographical detail (LA/UA identifiers), albeit at the expense of a more limited range of individual variables.

In addition, versions of the IL-SAR and SL-HSAR with more detailed geographical identifiers (UA/LA) referred to as the Controlled Access Micro-Data Samples (CAMS) are available at the ONS VML. Use of the CAMS data has relatively limited compared to the SARs due to the more restrictive arrangements surrounding access. The main factor that drives the structure, content and size of the SARS range of micro-data products is ensuring that the data is non-disclosive. The key principal with respect to sample size is outlined in GSS guidance for disclosure control:

Microdata based on a larger sample will have a greater absolute disclosure risk than a smaller sample…The larger the sample size in a set of microdata, the greater confidence an intruder can establish a possible identification, in that the larger sample size reduces the likelihood of a similar

individual existing outside the sample. Therefore microdata based on larger samples should be treated as more risky.

Sample size is clearly not the only measure used to maintain the anonymity of data. Restrictions in terms of geographical detail and the further suppression of data in relation to small groups (e.g. those respondents with extreme values for age) further reduce the risk of an individual being identified from these files. The range of data sets available via SARs highlight the trade-offs that can be made between sample size, suppression of data and methods of access in protecting the anonymity of Census respondents.

Following a consultation exercise, in March 2012 ONS published the 2011 Census Prospectus which sets out the release plans for 2011 Census statistics[43]. Included in this prospectus are proposals related to microdata files that will be made available in 2013. Three levels of micro-data are planned. A non-disclosive Public Use file will be available for download from the ONS website. This file will be less detailed than the file made available under EUL in 2001, with possibly no geographical identifiers and less detailed socio-economic classifications. Safeguarded files will sit somewhere between the EUL and Special licence files in terms of detail will also be available for download. Those requiring access would sign up to a 'special user agreement' with terms and conditions attached (although access would not require Approved Researcher status). Finally, files containing anonymised but identifiable data similar to the 2001 Controlled Access Microdata Sample (CAMS) will be held in the ONS VML facility. There are currently no plans to make Approved Research Census microdata files available in the Secure Data Service.

ONS Longitudinal Study

As with the Controlled Access Micro-Data Samples (CAMS), the ONS Longitudinal Study is only accessible onsite at ONS offices via the VML facility. User support is provided by CeLSIUS (see Chapter 2) with applications to use the ONS LS going through the Approved Researcher process outlined above. As well as the instances of the VML that exist within the offices of ONS within London, Titchfield and Newport, 2 further 'remote' instances of the VML (RVMLs) are located at NISRA and Scottish Government Offices located in Belfast and Glasgow respectively. These locations were developed prior to the launch of the Secure Data Service and were introduced in order to improve accessibility to the VML for Northern Ireland and Scottish based researchers. Typically, users of the RVMLs would first be required to visit the VML at a supervised ONS office to gain expertise and demonstrate their competence in the use of the system. Subject to the agreement of the VML team, researchers would then typically be allowed to access data via an RVML setting. It is of relevance to note that despite the availability of the technical infrastructure,

[43] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/2011-census-prospectus.pdf

the Longitudinal Study is not included in the range of data sets that can be accessed via the RVMLs[44]. In the case of NISRA, this is in spite of the RVML terminal being co-located with the safe setting operated by the NILS-RU and the expertise of staff that supervise that setting. The CAMs data is also excluded from the list of data sets that can be accessed via the RVMLs, pointing again to the unique conditions that surround access to the Census micro-data compared to other ONS sources.

3.5 Data Linking within ONS

The obligations of those responsible for the production of National Statistics that utilise data matching techniques are outlined in the National Statistics Code of Practice: Protocol on Data Matching[45]. The role of data matching in ONS and the principles to be adhered to are outlined in the document ONS Corporate Policy – Data Matching[46]. Particular emphasis is placed upon the procedures and standards that are required to implement statistical matching exercises involving ONS data. The scope of the document is important in understanding the nature of data matching that occurs within ONS. The Corporate Policy covers (bullets 3.1-3.3)

- Matching exercises where ONS supplies identifying data to other bodies (including contractors working for ONS) for matching or linking exercises.
- Matching exercises where ONS business areas allow non-ONS persons access to ONS datasets in order for them to match or link it, including where non-ONS datasets are to be integrated.
- Matching exercises in ONS using one or more data sources originating other than in ONS.

The scope of the document however does not include matching within ONS of sources originating wholly within ONS for wholly ONS purposes where it is felt that no significant governance or legal issues arise (bullet 3.4). The Office for National Statistics has previously linked to produce new statistical products and enhance the quality of existing resources in a cost effective manner, three examples of which are outlined below.

1. The Virtual Micro-data Laboratory was to set up primarily to provide access to business micro-data held by ONS. The most important source of business data has been the Annual Respondents Database, derived from responses to ONS's main business survey, the Annual Business Inquiry. The sampling frame for all ONS business surveys is the Inter-Departmental Business Register. All enterprises and establishments covered by ONS

[44] The VML Customer Request form is the form through which researchers provide project specific information that is not required on the AR form. This includes details of data sets required and the preferred location from which data will be accessed. http://www.ons.gov.uk/ons/about-ons/who-we-are/services/virtual-microdata-laboratory/accessing-the-vml/how-to-access-the-vml/customer-request-form.doc

[45] http://www.google.co.uk/url?q=http://www.ons.gov.uk/ons/guide-method/the-national-statistics-standard/code-of-practice/protocols/data-matching.pdf

[46] See section 5 of http://www.ons.gov.uk/ons/about-ons/who-we-are/services/unpublished-data/mrp-guide.doc.

business surveys therefore contain IDBR reference numbers. This has enabled, for example, information on business productivity from the Annual Respondents Database to be linked to information held on these businesses from other surveys, including research and development, e-commerce, capital expenditure, innovation and employment relations. This has enabled analysis to be carried out where collected data from a single source would not have been possible, thereby maximising the value that can be extracted from these business surveys (see Ritchie, 2008)[47].

2. Census data has also been used in data linking studies conducted within ONS. Jenkins (2008)[48] undertook an exercise to study the feasibility of linking ASHE and Census data for 3 geographical areas: Bedfordshire, Bexley and Cornwall based on matching gender, names and dates of birth. A common criticism of the ASHE data set is that it contains relatively limited information on personal characteristics (age, gender) and does not collect data on many characteristics that are known to be important determinants of earnings (e.g. ethnicity, disability, household composition). Match rates of between 50-75% were achieved, with rates being highest in Cornwall due to the high proportion of people who both lived and worked in the same area.
3. The Census has also played a pivotal role in the study of survey non-response. The 2001 Census Link study examined the determinants of response to 6 Government Surveys that were conducted around the same time as the 2001 Census (the Expenditure and Food Survey, the Family Resources Survey, the General Household Survey, the Omnibus Survey the National Travel Survey and the Labour Force Survey) by linking data on survey outcomes to Census data. In such studies, the Census provides a unique opportunity to examine the characteristics of those sampled households where contact could not be established and those households who refused to participate in the study (see Durrant and Steele, 2009)[49].

An important characteristic of these linking studies is that they are generally conducted on *sources originating wholly within ONS for wholly ONS purposes*[50]. Under such circumstances, there are no

[47] Ritchie F (2008), Secure Access to Confidential Micro-data: Four Years of the Virtual Micro-data Laboratory: Economic and Labour Market Review, vol. 2, no.5. Available at: http://www.ons.gov.uk/ons/rel/elmr/economic-and-labour-market-review/no--5--may-2008/economic---labour-market-review.pdf

[48] Jenkins J. (2008). Linking the Annual Survey of Hours and Earnings to the Census: a feasibility study. Economic and Labour Market Review, vol. 2(2) http://www.ons.gov.uk/ons/rel/elmr/economic-and-labour-market-review/no--2--february-2008/economic---labour-market-review.pdf

[49] Durrant G. and Steele F. (2009), Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. Journal of the Royal Statistical Society, vol 172(2).

[50] The data linking activity that has been undertaken within the Virtual Micro-data Laboratory has in some cases included data that was not originally collected by ONS. These data sets include the Workplace Employment Relations Survey (Department of Business Innovation and Skills) and the National Employers Skill Survey (Learning and Skills Council). Where external data has been brought in to the VML, the informed consent of respondents has been sort. See Aumeyr M, Davies R, (2009), 'Characteristics of Respondents Giving Informed Consent for Linking of Business

requirements to obtain the informed consent of respondents regarding the additional linking activity that could be undertaken on their data. For those linking activities based upon surveys where participation is voluntary (such as the Labour Force Survey), the standard survey pledge which informs respondents that their data will be used for statistical purposes is sufficient for allow in-house matching exercises to be undertaken. Whilst the upmost care is taken to maintain the security and confidentiality of respondents, usual data management policies and procedures apply.

3.6 Enhancing the England and Wales Longitudinal Study

The ONS Longitudinal Study is based on a 1% sample of the population of England and Wales. It would therefore be expected that some 30,000 residents of Wales would currently be LS members. ONS has recently completed the integration of 2011 Census data in to the LS and the new LS Database encompassing 2011 Census data should be available by the autumn of 2013, subject to the successful completion of testing. At this point, some LS members will have been included in the LS for a period of 40 years covering 5 waves of Census data. Due to its small sample size, the ONS Longitudinal Study has only been used on a relatively limited basis within a Welsh context. Exceptions to this include research conducted by the Welsh Language Board in relation to how individuals report their ability to speak Welsh from one Census to the next (Jones, 2005)[51] and the effects of outward migration from Wales on language use (Jones, 2012)[52]. An obvious barrier to using the ONS LS for Welsh level analysis is its relatively small sample size. Nonetheless, it remains the case that 30 thousand Welsh LS members represents a considerably larger number of people than those that are included in other longitudinal data sources, such as the British Household Panel Study and the Millennium Cohort Study. Elsewhere in the UK, the ONS LS has been used to study the issue of poverty and inward migration to Cornwall (Williams and Champion, 1998)[53] and examine migration into and out of the south east of England (Champion, 2011)[54]. An ongoing project (LS: 30112[55]) is looking at the effect of migration on labour market outcomes (the 'escalator effect') for the 8 'Core Cities Group' of England (Birmingham, Bristol, Leeds, Liverpool, Manchester, Newcastle, Nottingham and Sheffield). It can therefore not simply be assumed that the ONS LS is not suitable for Welsh level analysis. The feasibility of any Welsh level research will be context specific and the findings of the Beta Testing projects that are currently taking place and have been asked to examine the feasibility of Welsh level analysis (see Chapter 2) will be of interest.

Data: Analysis of the 2004 Workplace Employment Relationship Survey and the 2007 National Employer Skills Survey' *ONS Survey Methodology Bulletin* 64 pp 45-62.

[51] Jones H. (2005), Ability to Speak Welsh in the Censuses of Population: A Longitudinal Analysis. Population Trends vol. 122, Winter

[52] Jones H. (2012), A Statistical Overview of the Welsh Language. Welsh Language Board.

[53] Williams and Champion (1998), Cornwall, Poverty and In-Migration. Cornish Studies, Second Series (6): 118-126.

[54] Champion T. (2011), Testing the Return Migration Element of the 'Escalator region' Model: an analysis of migration into and out of South-East England. Cambridge Journal of Regions, Economy and Society.

[55] http://celsius.lshtm.ac.uk/projects/championEscalator.html

Such projects demonstrate that the ONS LS could provide a useful research resource for Wales. It may also be possible to integrate further data in to the ONS LS for LS members. External data has generally not been linked to the ONS LS due to difficulties surrounding the sharing of data between government departments. In theory, the linking of data which includes NHS patient id numbers to the ONS LS should be feasible. The HSCIC holds a look-up table containing the reference numbers of LS members against NHS identifiers. Therefore, subject to necessary approvals, it should be relatively straightforward to link additional data about individuals into the ONS LS if that data can be assigned an NHS reference number, or an anonymised version thereof. The coverage of such data could extend beyond the remit of health. For example, the allocation of NHS numbers to the National Pupil Database raises the possibility that detailed demographic and socio-economic information from the Census could be linked for 1% of pupils attending school in Wales. For example, such a linked data set would be able to provide information on absenteeism and attainment by parental social class rather than relying upon a free school meal status as a proxy for the economic conditions of the household. The effects of a range of household characteristics (e.g. family structure, housing conditions) upon different educational outcomes could potentially be considered.

3.7 Developing a Welsh Census Based Longitudinal Study

When established, the sampling fraction for the ONS LS was set at 1% in order to provide data on approximately half a million people, a sample size that was felt to be necessary to provide sufficient data to examine occupational mortality. This requirement subsequently shaped the development of the NILS, where a sampling fraction of 28% similarly provides data for approximately half a million people in Northern Ireland. At 5%, the Scottish sampling fraction is relatively small and reflected funding difficulties at the time that the SLS was established. Given that responsibility for policy in the area of health is devolved to the Welsh Government, a Census based Wales Longitudinal Study that provided comparable levels of statistical power would imply a sampling fraction of approximately 17%. As noted earlier, such an increased sampling fraction would result in a higher risk of disclosure compared to the existing ONS Longitudinal Study.

Where linking involves data being transferred in to or out of ONS, the standards and requirements of the Corporate Policy on data matching need to be met. These include issues surrounding governance (who is responsible for the matched data set, methods, data destruction), compliance with any survey pledges, ensuring the matching exercise is lawful, security (are necessary organisational and technical measures in place) and the establishment of an ethics committee to provide oversight. In the context of the emphasis placed upon maintaining the security and anonymity of Census responses, preliminary discussions with ONS indicate that it is unlikely that disclosive Census micro-data to be released to the Welsh Government for the purposes of

developing a Census based Welsh Longitudinal Study, although this would need to be confirmed. The Statistics and Registration Service Act did introduce significant reforms to the governance of official statistics, including provisions that allowed for government departments to apply for Information Sharing Orders (ISO) to set up Sharing Gateways for the purposes of producing 'Official Statistics'. However, these provisions require secondary legislation before such data sharing can occur. In practice their use has been limited and has tended to support flows of data in to ONS. Desai (2012) outlines a number of reasons why ISOs would not assist in the establishment of a Wales Longitudinal Study. Most significantly, as ONS is the statistical office for both England and Wales, any ISO for Wales must be laid before parliament by the ONS. Furthermore, ISOs do not compel government departments to share data, they are time consuming to prepare and relate to specific uses of single data sets (i.e. they do not establish long term relationships). The SRSA would therefore not appear to support the development of a Census-based Wales Longitudinal Study outside of ONS.

The Statistics and Registration Service Act does however also describe a number of functions that are to be undertaken by the Statistics Board. Whilst the emphasis upon maintaining the confidentiality of Census respondents remains paramount, the SRSA underlines the important role that ONS has in developing data resources and supporting research. It should be noted that functions associated with the provision of statistical services and the promotion of statistical research are expressed as *discretionary powers* in the Act as opposed to *duties* of the Board. Nonetheless, the statement of these powers within legislation are of relevance to the possible establishment of a Census based Wales Longitudinal Study with the collaboration of the ONS. Firstly, Section 22 'Statistical Services' states that:

> *The Board may provide statistical services to any person in any place within or outside the United Kingdom* and that these services may include *collecting, adapting and developing data.*

Secondly, Section 23 'Statistical Research' now explicitly states that one of the functions of the Board is to support research:

> *The Board may promote and assist statistical research, in particular by providing access (where it may lawfully do so) to data held by it.*

Whilst the SRSA suggests that ONS could in principle support the development of a Wales Longitudinal Study, issues surrounding resources, ownership and procedures would need to be addressed. The necessary technical and security infrastructure would need to be established at the HSCIC in Southport so that additional tracing of LS members within Census data could be undertaken. The creation of an enhanced version of the ONS Longitudinal Study would also require a Business Impact Assessment to be conducted within ONS that examined the risks

associated with creating the new data set. The Joint Information Assurance Policy Statement issued by the three UK Census Offices outlines the policy and commitments that these organisations had adopted on Information Assurance in relation to the 2011 Census. Given the scale, complexity and sensitivity of the Census, the statement asserts that *'the Census Offices are risk averse and are prepared to tolerate a low level of risk in the achievement of their business aims'*. The level of impact associated with a compromise in confidentiality associated with the LS has been undertaken by ONS using Business Impact Level tables (see HMG IA Standard No. 1 – Technical Risk Assessment). The research version of the ONS LS that is created and transferred to the server that be accessed by USOs for the purposes of supporting research access has been classified at Impact Level 3 'Restricted'[56]. The VML has been assessed as providing the necessary level of security to support research access to the LS. However, an increase in either the size of the LS or the inclusion of additional linked data sets in to the LS would require the level of Business Impact to be re-assessed. An increased impact level could impact upon the usability of the resource.

3.8 Concluding Comments

The establishment of Census based Longitudinal Studies by ONS, NRS and NISRA each relies on the fact that these statistical offices have responsibility for conducting and administering the Census. Whilst ONS is responsible for undertaking the Census in England and Wales, NRS and NISRA are the custodians of their Census data and have therefore been in a position to develop longitudinal studies in a way that has not been possible in Wales. The main area of research that is conducted using these Longitudinal Studies relates to public health. Given the devolution of powers to the Welsh Government in the field of Health and Health Services under Section 5 of the Government of Wales Act 2006[57], the inability to develop a longitudinal database that is capable of contributing to the evidence base in Wales to inform the decisions of policy makers in a way that has been achieved in Northern Ireland and Scotland would appear to be inequitable.

The discussion in this chapter has highlighted the upmost seriousness with which ONS is committed to maintaining the confidentiality of Census respondents. This commitment manifests itself via publically issued assurances given to Census respondents and the more restrictive access arrangements associated with Census micro-data compared to other sources of survey data. Under such conditions, it is unlikely that ONS would release disclosive Census data to the Welsh Government for the purposes of developing a Census based Wales Longitudinal Study. The development of a Census based Wales Longitudinal Study therefore implies either the Welsh

[56] ONS staff responsible for the development of the LS actually work with a more disclosive version of the database that is classified at Impact Level 4.

[57] http://www.assemblywales.org/bus-home/bus-legislation/bus-legislation-guidance/bus-legislation-guidance-documents/legislation_fields/schedule-5.htm

Government becoming responsible for undertaking the Census in Wales or attaining ownership of Census data collected on its behalf or the Welsh Government collaborating with ONS in developing a Census based Wales Longitudinal Study.

At the time of writing, no micro-data research resources arising out of the 2011 Census are yet available to the research community. In terms of developing a Census based analytical resource within the short to medium term, collaboration with ONS would need to be sought. However, whilst the SRSA underlines the important role that ONS has in developing data resources and supporting research, significant issues would need to be addressed in terms of resources, governance and infrastructure necessary for the development of a resource that may have a higher Business Impact Level than the existing England and Wales Longitudinal Study.

There are two caveats to this approach. Firstly, WG have made considerable investments in the SAIL data linking facility based at Swansea University. Whilst it would be envisaged that SAIL (and the NHS Welsh Informatics Service) would have an important relationship with an ONS based Wales Longitudinal Study (e.g. as possible suppliers of suitably anonymised but linkable Welsh data), the investment in another data linking facility based at ONS may not be feasible. Secondly, the SRSA does not impose a duty on ONS to develop new statistical resources and ONS may simply not have an 'appetite' to support the development of such a resource. A sensible (and timely) approach in the short term may be to investigate the feasibility of linking sources of Welsh administrative to the England and Wales Longitudinal Study. Clearly, as a 1% sample this study would not be suitable for the study of small population sub-groups in Wales. However, with an estimated 30,000 Welsh LS members, it remains a potentially important source of longitudinal data for Wales. Attempting to develop links to a large administrative database such as the National Pupil Database would appear to represent a sensible first step. The results of such a project could then inform discussions as to whether or not it would be worthwhile for the Welsh Government to invest in the construction of a larger Census based longitudinal study for Wales.

Chapter 4: Non-Census Data and the Wales Longitudinal Study

4.1 Introduction[58]

The three UK Longitudinal Studies are similar in their design, being based on collections of data from routine administrative sources such as the Census and registration systems. The section of individuals for inclusion within the each of these studies is based upon the day and month of birth. However, there are differences in how the populations for these studies are selected. While the ONS-LS uses four birthdates, and the SLS is based on 20 birthdates, the NILS includes individuals on the basis of 104 birthdates. For both the ONS Longitudinal Study and the Scottish Longitudinal Study use Census population data linked to the National Health Service Central Register (NHSCR) to define their cohorts. In Northern Ireland, the Health Card Registration system data (a database of all people registered for health services in Northern Ireland including name, date of birth, sex and address) is used as the core of the NILS sample. The sample for the NILS was selected initially from the Northern Ireland Health Card Registration System on the day of the 2001 Census. The NILS members were then matched to the 2001 Census records[59].

The main disadvantage of the Northern Ireland approach arises from the problem of list inflation, a situation where health records continue to exist on the registration system for individuals who no longer exist in Northern Ireland. Estimates published by the NILS-RU indicate that the number of patients registered on the Health Card Registration System exceeded the Census based figure for the Northern Ireland population by 4.7%. However, the benefit of this approach is the frequency with which the NILS database can be updated. Although only limited information is available from administrative data, timelier (and arguably more accurate) annual population counts of registered people by age and gender can be produced. Irrespective of the method of construction, it is noted that for each of the three UK Longitudinal Studies Census data plays a central role in providing data related to socio-economic characteristics, housing conditions and family structure. For many research projects, the availability of information derived from the Census would determine eligibility for inclusion within samples for analysis.

The discussion of so far has focussed on issues surrounding the construction of a Census based Wales Longitudinal Study. The ability to draw upon this source in the construction of a Wales Longitudinal Study may be limited due to ONS having responsibility for the Census in England and

[58] This section incorporates data from the Annual Population Survey which is produced by the ONS and is accesses via special licence from the UK Data Archive, University of Essex, Colchester. None of these organisations bears any responsibility for the analysis or interpretation undertaken here.

[59] See Johnstone, F, Rosato, M. and Catney G. (2010) The Northern Ireland Longitudinal Study: An Introduction, NILS Working Paper 1; available at: http://www.qub.ac.uk/research-centres/NILSResearchSupportUnit/FileStore/Filetoupload,238885,en.pdf

Wales. Within the SAIL system the NHS Administrative Register (the Welsh equivalent of the NHCSR) has formed the basis of the population spine through which data sets can be linked (see Lyons et al 2009)[60]. The NHSAR contained the administrative details of all individuals who have registered or accessed health services in Wales (primarily via registrations with GPs). All persons registered with the National Health Service (NHS) in England and Wales are assigned a unique 10-digit NHS number, and this is used as the personal identifier for patients across different NHS organisations. From August 2009, responsibility for the management of administrative information (demographic data) for NHS patients within Wales became the responsibility of the Welsh Demographic Service. Given its central role in linking, it is therefore important to be aware of the possible limitations of NHS administrative information as a source of demographic data.

The second key issue that emerges in the absence of Census data is what sources of socio-economic data relating to the population of Wales can be incorporated in to the WLS? The demographic details maintained by WDS are limited, including name, address, date of birth, gender and General Practice. At the outset, it is acknowledged that Welsh Government Surveys such as the National Survey and the Welsh Health Survey have included consent to link questions for the purpose of data linking. These surveys could therefore be used to contribute socio-economic information about their respondents to the WLS. However, these surveys are relatively small cross sectional surveys and difficulties may arise in trying to combine data sets based upon surveys with different methodologies or where responses to questions are not coded consistently. The largest single source of survey data collected in Wales is the Annual Population Survey and would be the preferred alternative source of socio-economic for inclusion within the WLS. Broadly speaking, the APS in Wales encompasses pooled responses from four successive quarters of the Labour Force Survey and a Welsh Government funded boost to the LFS (the Scottish Government also fund a boost to the LFS in Scotland). The second part of this chapter provides an overview of the APS, particularly in relation to its longitudinal properties, and considers whether this source of data could be integrated in to the Wales Longitudinal Survey.

4.2 NHS Administrative Records as a Population Spine

Insights from NHS Patient Registers

Official statistics related to information about the populations registered with GP practices at Strategic Health Authority (SHA) and Primary Care Organisation (PCO) level for England and Wales are produced by the HSCIC. These figures are derived from the NHS Patient Register; a database of all persons registered with a GP that is maintained primarily for the purpose of transferring medical records and for managing payments to GPs. It is noted that the NHS Patient Register is not synonymous with the NHS Central Register discussed above. The NHS Patient

[60] http://www.adls.ac.uk/wp-content/uploads/2011/08/SAIL-Building-a-National-Databank.pdf

Register is a complete list of all persons registered with a GP in England and Wales whereas the NHSCR is a centralised system for patient administration, which includes weekly updates for births/deaths, new registrations with GPS and other administrative changes (e.g. name changes, enlistment to the armed forces, cancer registrations).

Cross national comparisons of the relative size of the GP populations are published within ONS Regional Trends, with the latest data relating to 2008[61]. Table 1 compares GP population estimates with ONS mid-year population estimates based upon 2001 Census data. It is not possible to subtract data on 'Special Populations' from published mid-year population estimates and therefore comparisons between the 2 sources can only provide an indication of the scale with which the GP population exceeds office Census based estimates of the UK population. Overall, it is observed that the GP population exceeds the Census based population by approximately 5%. Within Wales, the GP population for 2008 is estimated to be approximately 124 thousand more than the Census based estimate.

Table 1: UK Comparisons of ONS and GP Population Estimates: 2008

	GP Population	**ONS Population**	**GP/ONS Differential**
England	53939900	51464600	4.8%
Wales	3113700	2990100	4.1%
Scotland	5473200	5168500	5.9%
Northern Ireland	1848300	1775000	4.1%
	64375100	61398200	4.8%

Sources: ONS Regional Trends, 2009; ONS Mid-Year Population Estimates

Within Wales, the Welsh Government produces an annual publication that provides workforce data relating to General Medical Practitioners[62]. The statistical release presents data on the number of practitioners, patients per practitioner and the demographic characteristics of practitioners. Information on the number of registered patients per practitioner and the number of practitioners can be used to provide an estimate of the GP population in Wales. These are presented in Figure 1 which again provides comparisons with ONS Mid-Year Population Estimates. It can be seen that within Wales, the GP based measure of population is consistently higher than ONS estimates. Furthermore, there is some indication to suggest that the size of this differential has widened between 2001 and 2010. It is noted however that Mid-Year population estimates are based upon the latest available Census data (2001) combined with annual estimates of population change arising from births, deaths and migration and errors in these estimates of population change will be

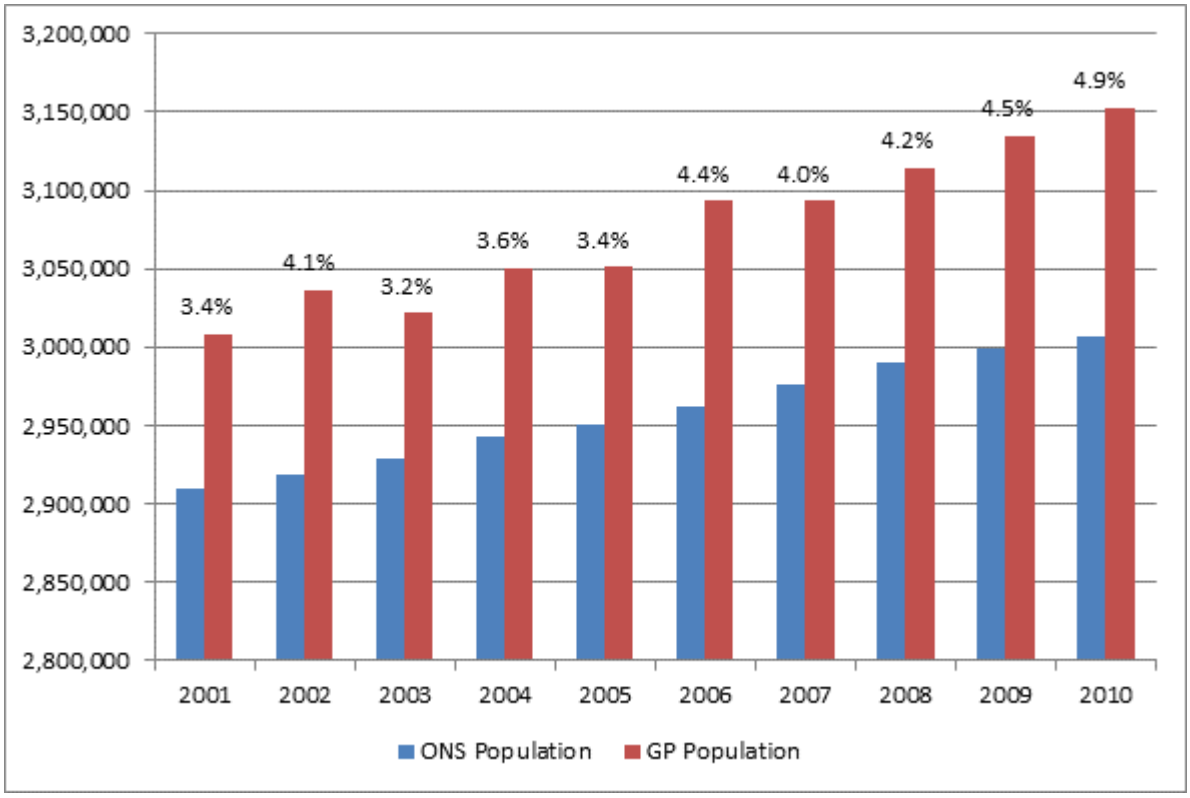
[61] The HSCIC also produce estimates of GP registered populations scaled to ONS population estimates. See: http://www.hscic.gov.uk/searchcatalogue?productid=4710&q=title%3a%22Attribution+data+set+GP+registered+populations%22&sort=Relevance&size=10&page=1#top

[62] http://wales.gov.uk/docs/statistics/2011/110322sdr442011en.pdf

compounded over time. Both GP workforce data and ONS population estimates are also provided for Local Health Boards. Comparisons of this data in Figure 2 also reveal that the scale of this differential varies across Wales. In both relative and absolute terms, the size of this differential is largest within the Abertawe Bro Morgannwg and Aneurin Bevan Health Boards. In each case, the size of the GP population is 37 thousand larger than that provided by ONS estimates.

The discrepancy between population estimates based on Census figures and the registered lists of GPs is commonly referred to as list inflation. There are several reasons why some patients on GP lists no longer exist, including death, emigration, migrants returning home or individuals moving house. List inflation has been shown to vary by age, gender and area characteristics. It is noted to be particularly problematic among young males living in relatively deprived inner city areas[63]. Non-existent patients are often termed 'ghosts'. To avoid miss-allocation of resources, administrative processes are in place for people moving between UK territories to ensure that their medical records are up to date. Since 2004, a regular National Duplicate Registration Initiative has been run by the National Audit Office to detect inaccuracies in GP lists that could distort the allocation of funds[64].

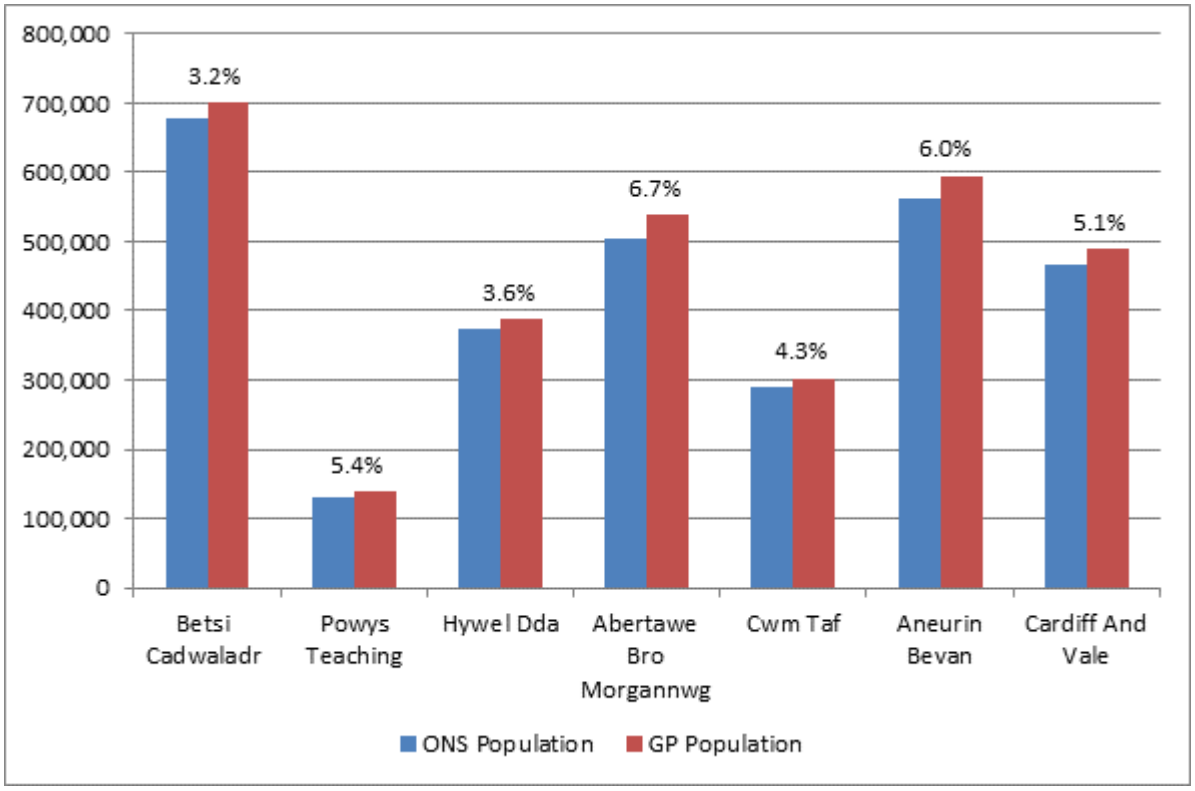
Figure 1: GP and ONS Population Trends in Wales



Sources: General Medical Practitioners 2011, Welsh Government; Mid Year Population Estimates, ONS

[63] See Ashworth et al, 2005 available at: http://fampra.oxfordjournals.org/content/22/5/529.full.pdf+html
[64] http://www.audit-commission.gov.uk/sitecollectiondocuments/downloads/ndrireport2012.pdf

Figure 2: Regional Variations in GP and ONS Population Estimates in Wales



Sources: General Medical Practitioners 2011, Welsh Government; Mid Year Population Estimates, ONS

Insights from the ONS Migration Statistics Improvement Programme

Patient register information has always been a key source of data used in the construction of population estimates produced by ONS. The main use of this information has been for estimating internal migration between local authorities. Previous studies have been undertaken by ONS in order to consider the relative accuracy of patient register information as a source of demographic data. However, the importance of administrative data as a source of demographic data has come to the fore recently in relation to the Beyond 2011 programme. As part of its Migration Statistics Improvement Programme, in 2012 ONS published the report *Using Administrative Data to Set Plausibility Ranges for Population Estimates.*[65] The report investigates the feasibility of developing ranges from different administrative sources within which population estimates may be expected to fall. The report presents the results of analysis that has initially been undertaken on children up to the age of 15. This choice of this age group reflects the importance attached to achieving an accurate 'base' in the context of the cohort component methodology used in population estimates; the greater quantity and quality of administrative data for this age group; i.e. a greater tendency for them to be registered with a GP; the (current) existence of universal child benefit and the

[65] http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/using-administrative-data-to-set-plausibility-ranges/index.html

accessibility to ONS of both aggregate and record level data to School Census data for this age group.

Within this age group a variety of data sources are utilised to produce lower and upper limits of population estimates, including registration of births (used for those under the age of 1), child benefit receipt, School Census and the Patient Register. The analysis reveals that patient register data generally forms the basis of the derived upper limits for population estimates. Those under the age of 1 are the exception to this, where the registration of births generally provides the upper limit for population estimates. Among older age groups, Patient Register counts exceed Child Benefit counts across most local authority areas and tend to form the basis of upper limit estimates. Lower bound estimates are typically derived from the School Census, which on average accounts for 92% of children (pupils in the non-maintained sector are not covered by this source). Due to the completeness (and possible inflation) of population estimates based upon the Patient Register, the report provides detailed comparisons between Patient Register data and Census based estimates for children up to the age of 15.

Full results of comparisons between GP and ONS based population estimates are presented in Annex 1. It can be seen that the degree of comparability between the 2 data sources is much higher among younger children reflecting the quality of Patient Registration data for this group. Among older age groups, the estimated gap between the 2 sources widens. Older children are more susceptible to errors associated with delays in parents re-registering with a doctor following a change in address. Figures 3 and 4 present comparisons between mid-year population estimate and Patient Register data for children aged 12-15, the age group for whom discrepancies between the 2 sources are widest. It can be seen that differentials between GP and ONS based population estimates are relatively large in Cardiff and Swansea, possibly reflecting the relatively dynamic nature of the population within the 2 largest cities in Wales. A further issue affecting this group is related to those attending boarding school who may appear either at the parents' address or the boarding address on the Patient Register. Although present across all parts of Wales, the relative concentration of such schools in Conwy and Denbighshire may contribute to relatively high numbers of children appearing on the patient register in these areas. The presence of Armed Forces bases which provide medical facilities for families may also contribute to differences between the 2 population sources.

ONS are continuing to develop this area of work for other age groups within the adult population. Although results are not available, a number of issues of relevance are identified. Among the young adult population, the Patient Register produces lower population estimates in University based towns. ONS are therefore attempt to adjust the Patient Register with data on student records from the Higher Education Statistics Agency by essentially moving students to where they are studying based upon their term time address. The quality of the Patient Register for the

working age population and the over-retirement population is not considered to be as good as that for children. Among the working age population (24-59/64), Patient Register and ONS estimates align more closely among older age groups. Differences between the sources are also smaller for people over retirement age due to the increased reliance of this group upon health services.

Figure 3: ONS and Patient Register Population Estimates: Ages 12-15, Males

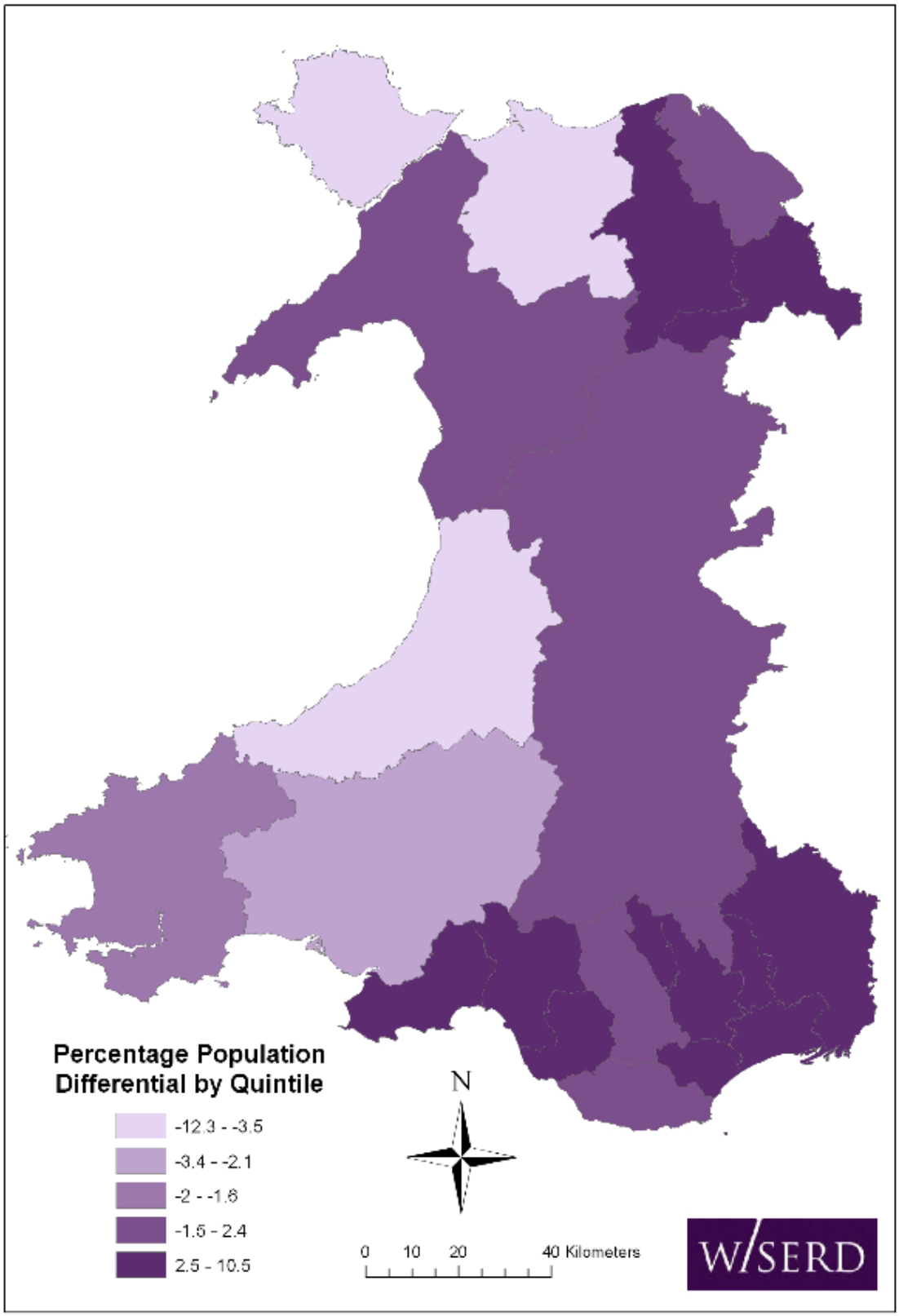
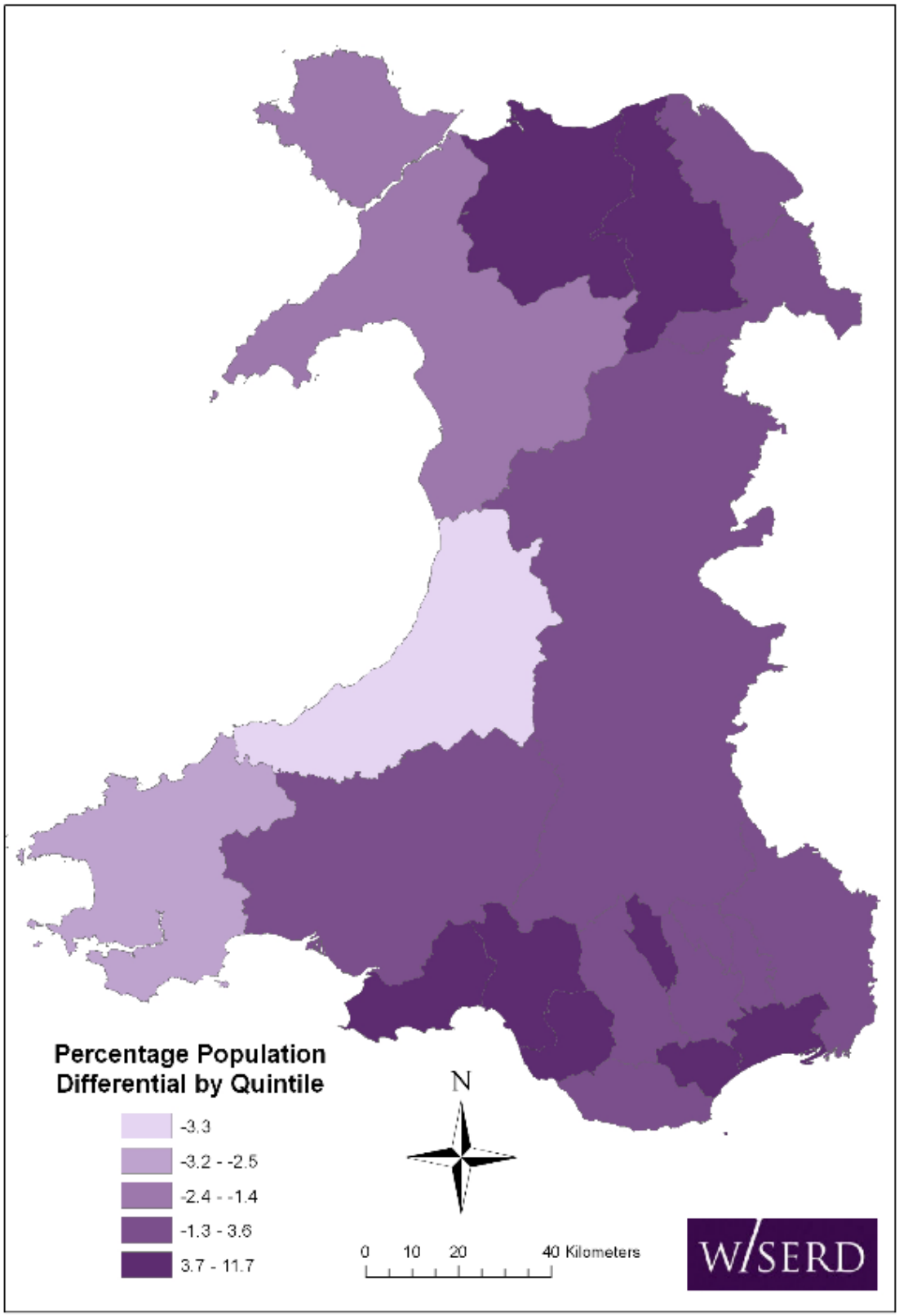


Figure 4: ONS and Patient Register Population Estimates: Ages 12-15, Females



Insights from the ONS Beyond 2011 Programme

The most recently available research that has reviewed the scope and quality of the NHS administrative data as a source of population data has been conducted by ONS as part of its Beyond 2011 Programme[66]. The report presents the most up to date picture regarding the completeness of the NHS Patient Register, based upon a comparison made with 2011 Census data with a snapshot of the NHS Patient Register taken within a week of the Census reference date. The findings generally concur with the results outlined above. The NHS Patient Register provides broad coverage of people within England and Wales; at national level the Patient Register is demonstrated to exceed the Census 2011 estimated by 4.3%. Three quarters of local authorities were within five per cent of census estimates, although local area variation does occur when data are compared locally by both age and sex, highlighting that overall differences represent the combined effects of 'different issues which impact in different ways locally'. The majority of local authorities that show higher NHS Patient Register counts are characterised as being large urban areas, with eight London boroughs being within the top ten local authorities with the highest percentage difference. There are only a small number of local authorities that show a lower count in their 2011 NHS Patient Register than their Census estimates. Such areas are often characterised by the presence of bases for the Armed Forces. Within Wales the NHS Patient Register count is greater than the 2011 Census estimates by 3.1 per cent, possibly reflecting the relatively lower proportion of people living in urban areas within Wales.

Insights from the ONS Longitudinal Study

The Migration Statistics Improvement Programme has sought to compare aggregate estimates of the population based upon administrative records have been compared with those derived from Census data. In terms of Patient Register data, 'list inflation' is clearly important in understanding why the numbers registered with GPs exceed Census based population estimates. However, certain groups are also less likely to respond to the Census. For 2011, the all person response rate to the Census in Wales was estimated to be 93% (94% overall for England and Wales), with young adult males (those aged 20-34) exhibiting the lowest levels of response at approximately 85%[67]. Analysis has been undertaken on the ONS Longitudinal Study (Smallwood and Lynch, 2010[68]) to compare differences in area of usual residence when comparing LS members from the 2001 Census with data from the National Health Service Central Register (NHSCR) as held on the

[66] Beyond 2011 Administrative Data Sources Report: NHS Patient Register, available at: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/news/reports-and-publications/sources/index.html**.** This is the first of a series of reports that will look at the value of different administrative sources for the purposes of the Beyond 2011 programme, including electoral data, schools data, university data and DWP/HMRC administrative records.

[67] http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/response-rates-in-the-2011-census.pdf

[68] **http://www.ons.gov.uk/ons/rel/population-trends-rd/population-trends/no--141--autumn-2010/index.html**

same day as the 2001 Census. This analysis provides an insight in to both the numbers of potential LS members who have not been linked to a Census record and the number of LS members who, on the date of the Census, were enumerated at a different address to that recorded by NHS administrative records. The analysis distinguishes the following 5 groups.

- Those in the same area on the Census and NHSCR (95.7%)
- Those on the 2001 Census but who are in a different area according to the NHSCR (2.8%)
- Those on the 2001 Census but whose NHSCR record had been cancelled prior to Census day (0.8%)
- Those on the 2001 Census but who only appear on the NHSCR later (0.7%)
- Those not on the Census but who are on the NHSCR at Census time.

In the case of the last category, 15% of people identified by the NHSCR as potential LS members do not appear on the Census. This group is difficult to interpret and may arise for a number of reasons including Census non-response, LS members who could not be linked or those who have left the country without informing their GP. Further analysis reveals that almost half of those registered at a different area on the Patient Register (Category 2) or who are not registered on the Patient Register (Category 3), eventually become registered as being in the same area as their Census return, highlighting the importance of delays in the registration system. London, the South East and the East of England have the highest numbers of LS members for whom the NHSCR places these people in a different Government Office Region (GOR) compared to the Census, reflecting the high levels of inward migration to this part of the UK. Males and those aged 20-39 are more likely to be either in a different place, absent from the Patient Register or absent from the 2001 Census. The analysis concludes that NHS administrative data is a good quality administrative source and in most cases matches with the information collected at Census.

4.3 The Annual Population Survey as a Longitudinal Research Resource for Wales

The important characteristic of the three UK Longitudinal Studies is the inclusion of Census data which provides background socio-economic data on people included in these studies. It is Census data that has allowed various health outcomes to be compared for different sub-groups of the population. If Census data cannot be incorporated into the WLS, what sources of data can be used to build socio-economic data in to a WLS? A number of options may be available. Firstly, administrative data sets could be used to provide socio-economic data. For example, data included in the National Pupil Database includes information on entitlement to Free School Meals. Such information could possibly provide a useful 'proxy' of the material conditions under which households with children attending state schools live. However, clearly no information can be developed for those households with no school aged children or children attending non-state maintained schools. The richest administrative sources of administrative data relating to the

material conditions of households are records held by the Department for Work and Pensions (DWP) in relation to benefit entitlement and tax information held by Her Majesty's Revenue and Customs (HMRC)[69]. Whilst the work of the Administrative Data Taskforce is exploring how administrative data can be made more accessible for the purposes of research (see Chapter 1), such data would still not contain the necessary occupational and employment relations information necessary to derive measures of social class, such as the National Statistics Socio-economic Classification.

Welsh Government Surveys such as the National Survey and the Welsh Health Survey have included consent to link questions for the purpose of data linking. Although relatively small cross-sectional surveys, it is conceivable that these sources could used to contribute socio-economic information about their respondents to the WLS. Pooling data collected from different sources can however be problematic in practice. The *Strategic Framework for Welsh Government Surveys, 2012-2017*[70] provides guiding principles for the conduct of surveys by the Welsh Government, including adherence to national standards[71] for the collection of key data items, including questions necessary for the derivation of social class. However, inconsistencies between surveys still emerge. For example, whilst the Welsh Health Survey measures the social class position of a household reference person who is identified as being the highest income householder[72] (the preferred method for allocating social class to a household), the National Survey for Wales records the social class position of the respondent[73]. Whilst both surveys use the correct harmonised questions, merging information on socio-economic status from these 2 sources would be problematic, particularly in relation to understanding the position and life chances of women. The Scottish Government has introduced a set of core and harmonised questions in its surveys as part of its *Long Term Strategy for Population Surveys in Scotland 2009-2019* for the purpose of pooling data across surveys in order to support so that *each survey can be combined to produce a large enough sample size for robust estimates of rarely occurring characteristics at a national level and of other characteristics at a small area level*[74].

The preferred approach to collecting socio-economic information about the population of Wales would be through the use of a single large source of survey data. The Labour Force Survey (LFS)

[69] See http://www.justice.gov.uk/downloads/statistics/mojstats/offending-employment-benefits-emerging-findings-1111.pdf for an example of how administrative data held by DWP and HMRC has been used to sutudy the relationship between offending, employment and benefits.
[70] http://wales.gov.uk/topics/statistics/about/survey/?lang=en
[71] http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-andquestions/index.html
[72] See section 5.2.2 in http://wales.gov.uk/docs/statistics/2012/120919technicalreporten.pdf
[73] See questions 101-116 in http://wales.gov.uk/docs/caecd/research/121025nsw2012AnnexBEnglishlanguagequestionnaire.pdf
[74] Details for the long term strategy for surveys on Scotland is available at: http://www.scotland.gov.uk/Topics/Statistics/About/SurveyStrategy

is a survey of households living at private addresses (excluding communal establishments) in the UK. Its purpose is to provide information on the UK labour market which can then be used to develop, manage, evaluate and report on labour market policies. The quarterly LFS launched in 1992 in GB and in 1994 in NI operated on a seasonal quarter basis: March-May (Spring), June-August (Summer), September-November (Autumn) and December-February (Winter). Following EU requirements, in May 2006 the LFS moved to calendar quarters (CQ's). The LFS collects information on personal characteristics, household structure, economic activity, health, education and training and earnings. Among those in employment, detailed information is collected on jobs held including occupation, hours worked, earnings and contractual status. The LFS achieves interviews with some 53,000 households per quarter. Information is collected on over 140,000 individuals, of which approximately 60% are of working age making it the largest regular household survey conducted in the UK. The LFS is designed to provide information on the labour market characteristics of a nationally representative sample of the UK population. Nonetheless, it is designed to provide robust labour market information at a national level and its sample size is insufficient to provide reliable data at local level. Therefore, annual datasets are produced to support local area analysis, combining information from the quarterly datasets with additional boost surveys.

The Annual Population Survey is a database derived from the Labour Force Survey that has been made available from the ONS on an annual basis from 2004. The APS data contains observations from three sources: the Quarterly Labour Force Survey (QLFS), a specific APS boost and the Local Labour Force Survey (LLFS). The LLFS was introduced separately in England (from 2000), Wales (from 2001) and Scotland (from 2003) and was designed to enhance or 'boost' the number of observations from the QLFS to provide more robust information at the local authority level. Although data derived from the LLFS is available prior to 2004 in an aggregated form (e.g. via NOMIS and the UK Data Service), the APS is the only format through which individual level responses from the LLFS is made available for research access. Whilst the information contained within the LLFS is based on the same survey questions as the more widely utilised Quarterly LFS there is one critical difference; addresses sampled as part of the LLFS are selected for inclusion on the basis of a rotational four year panel. The APS therefore provides the opportunity to track individuals in participating households for up to a period of 4 years.

Due to the rotational design of the survey (responses being carried forward to subsequent waves), considerable attention is given to ensuring the longitudinal integrity of the APS. For many questions the responses provided by individuals during the previous Wave are available to the interviewer so that the previous circumstances of respondents can be referred to when certain questions are asked. The APS data therefore contain a number of system variables that allow individuals and households to be uniquely identified. A matching exercise has been performed

based on these system variables to link information available for the same individuals over time. The APS data files contain data on individuals collected from both the main LFS survey and the Local Labour Force Survey. In the case of the LFS, a respondent may appear in the APS data files on two occasions one year apart. This corresponds to the interviews they provided in first and fifth Waves of the LFS. In the case of the APS, a respondent may appear in the APS data files on four occasions. This relates to households selected for inclusion in to the APS being interviewed once a year over a period of 4 years.

The construction and statistical properties of the longitudinal APS data set are outlined in the report *Unlocking the Potential of the Welsh Local Labour Force Survey: An Investigation into Labour Market Transitions in Wales* (Jones et al, forthcoming). The patterns with which individuals can appear within the APS are complex and are presented in detail in Annex 1. For ease of exposition, Table 4.1 provides summary information on the contents of the panel database. It can be seen that the panel database contains information on 1.5 million individuals for the period 2004 to 2010. Of these, approximately 10% (151,000) are from Wales, highlighting the relative importance of the Welsh Government funded boost to the LFS. It can be seen that the size of the boost in Wales is such that the APS sample in Wales is comparable to that achieved in Scotland. Within the UK context, Wales can claim to have the 'best' LFS data in terms of the relative sample size available within the APS.

Longitudinal data is not held of all respondents within the APS data. Within Wales, the APS database only provides a single year's worth of data for approximately 72,100 individuals. There are 2 groups of respondents who account for a relatively large proportion of `single appearances'; respondents who appear in the sample of the APS for the last time in 2004 and those who appear for the first time in 2010. Many of those individuals who appear in the panel for the first time in 2010 would be expected to provide either Wave 5 main sample interviews in the following year or appear in subsequent years of the Annual LLFS. Similarly, many of those individuals who appear in the panel for the last time in 2004 would be expected to have provided Wave 1 main sample interviews during 2003 or to have appeared in previous waves of the WLLFS[75]. Furthermore, given that it is addresses that are sampled for inclusion in to the LFS, some households who appear at an address (or individuals who appear within a household) are unable to provide a complete profile of longitudinal data over the period for which their address was sampled for inclusion in to the LFS. The relatively high proportion of respondents providing only a single year's worth of data should therefore not be regarded as being indicative of levels of attrition within the surveys that contribute to the APS database.

[75] The WLLFS predates the availability of the APS data sets from 2004. However, micro-data from the WLLFS has only been made generally available to researchers following the introduction of the APS.

Among those for whom longitudinal data is available (approximately 78,800), around 60% (46,900) only provide 2 periods worth of data. By definition, those who responded to the main LFS sample are only able to provide 2 years worth of longitudinal data. Among respondents from the enhanced sample, approximately 17,400 individuals have data spanning a period of 3 years, with a further 17,800 having data over a period of 4 years. However, it is noted that duration of data simply measures the time elapsed between respondents first and last appearing in the APS. These results do not imply that these individuals complete profiles of data that cover each intervening year. Therefore, whilst approximately 17,800 of respondents in Wales have data spanning a period covering 4 years, approximately 14,600 respondents actually provide 4 years worth of data.

To put these figures in to context, one of the main sources of panel data available in the UK which tracks individuals and households over time is the British Household Panel Survey. The first wave, 1991, contained information on approximately 5,500 households and interviewed 10,300 adults. A major development at Wave 9 (1999) was the recruitment of two additional samples to the BHPS in Scotland and Wales. Within Wales, this had the effect of increasing the size of the Welsh sample from 572 in 1998 to more than 3000 in 1999. During 2008 (the last year of BHPS data collected before the study sample was subsumed in to the new Understanding Society study), the Welsh sample size was approximately 2,500. Whilst the APS is not a dedicated panel data set that attempts to follow up individuals who move home, the analysis has demonstrated that it is a significantly larger source of longitudinal data for Wales than the BHPS or Understanding Society.

Table 4.1: Summary of Years Present and Duration within the APS (2004/2010)

	Main LFS Sample				**Enhanced Sample**			**Total**
	England	**Wales**	**Scotland**	**Northern Ireland**	**England**	**Wales**	**Scotland**	
Years Present								
1	56.0%	55.2%	55.7%	55.0%	54.9%	44.7%	41.4%	53.7%
2	44.0%	44.8%	44.3%	45.0%	21.6%	25.4%	25.1%	35.6%
3	0.0%	0.0%	0.0%	0.0%	12.1%	16.3%	18.1%	5.6%
4	0.0%	0.0%	0.0%	0.0%	11.4%	13.7%	15.3%	5.1%
Total	743,872	44,223	78,297	36,181	385,381	106,608	116,945	1,511,507
Duration (years present>1)								
2	100.0%	100.0%	100.0%	100.0%	42.1%	40.9%	37.4%	74.5%
3					26.2%	28.9%	30.5%	11.9%
4					31.7%	30.2%	32.1%	13.6%
Total	326,998	19,822	34,691	16,286	173,917	58,988	68,541	699,243

It is acknowledged that the longitudinal APS database does have limitations. Firstly, it is only possible to study transitions over a relatively short time period. Secondly, it is not a proper panel database in the sense that it is addresses and not individuals who are sampled for inclusion in to the constituent surveys. Those families who move house or children who leave home are not followed up. The panel data base derived from the APS will therefore under-represent some groups of more mobile individuals such as those who are younger, students and those of non-white descent. Thirdly, as a brand new data resource developed for the purpose of multivariate analysis, the characteristics of the sample are not clear in terms of attrition. No sample weights have been developed for this data set. However, the main advantage of the data set is that it is possible to analyse transitions among sub-groups of the population that are unlikely to be possible with the BHPS. This is illustrated in the exemplar studies included within the report of the longitudinal APS which look at employment transitions among the disabled and the young. The longitudinal APS database however retains many of the key advantages of the LFS, namely its extensive coverage of issues relating to education and employment, the comparability of information collected across time and the significantly larger sample size than that available from the BHPS described above and can provide a valuable source of longitudinal data for Wales.

The size of the WG funded boost to the LFS means that the APS is – with the exception of the Census – the single largest source of survey data regarding the Welsh population. Due to its sampling design, it is also arguably the largest single source of longitudinal data in Wales. Ideally, the APS would form an integral part of any Welsh Longitudinal Study. However, whilst the WG fund the Welsh boost to the APS, ONS remain the data custodians and are therefore responsible for maintaining the anonymity of respondents. Similar difficulties therefore emerge in the use of the APS for a Wales Longitudinal Study as those that were outlined for the Census in the previous chapter. The use of the APS data for the purpose of data linking by organisations outside of ONS would require the informed consent of respondents. Advice received from ONS indicates that this would need to be signed consent. Such consent to link questions would only be asked (as is the case with other follow-up questions) at the end of final wave of interviews to avoid and potential negative impact of such a question upon subsequent waves.

Several problems emerge in gaining written consent from respondents during their final LFS or APS interview. Firstly, whilst a large majority of Wave 1 interviews are conducted face to face, a majority of final wave interviews are conducted by telephone. For 2010 within Wales, it is estimated that only 18% of QLFS Wave 5 interviews and 21% of APS Wave 4 interviews are conducted face to face. A majority of consent forms would therefore need to be sent out after the final interview is conducted. The complex patterns of response to the LFS and APS presented in Annex 1 also indicate that many households fail to complete their 'full set' of interviews. Consent to link would therefore be being asked of the most attrition diminished sample. Problems in gaining

consent would be further hindered by the high proportion of responses (approximately a third) that are collected by proxy respondent (typically the spouse or partner of the intended respondent). Consent to link cannot be provided by proxy, making it difficult to gain consent to link from other household members. Finally, it is also noted that despite the particular value that the APS has within a Welsh context, it would be difficult to implement such changes in what is a WG funded boost to a UK wide survey. The inclusion of consent to link questions within LFS and APS interviews would therefore appear to be much more problematic than has been the case in other surveys and it is therefore likely that levels of consent achieved would be far lower than those achieved in other household surveys.

4.4 Concluding Comments

Both the ONS Longitudinal Study and the Scottish Longitudinal Study use Census population data linked to the National Health Service Central Register as the population spines from which participants in these studies are selected. In contrast, in Northern Ireland the Health Card Registration system data is used as the core of the NILS sample. If the Wales Longitudinal Study were to be established with the cooperation of ONS and build upon existing mechanisms in order for Census data to be incorporated in to the database, then it would be expected that the Census would form the basis of a population spine. While the case could be made for a 17% sampling fraction in Wales to achieve parity with existing Longitudinal Studies, anything larger would be out of line with the attitudes that prevail among those responsible for the three existing UK Longitudinal Studies.

If such a database were constructed via SAIL, the population spine would be based upon the NHS Administrative Register. The advantage of such an approach would be the ability to produce a database that provides a more dynamic and timely picture of the population of Wales. Whilst administrative health data does suffer problems associated with 'list inflation', the extent of such problems is not large. Analysis conducted for the Beyond 2011 Programme concludes that the NHS Patient Register provides broad coverage of people within England and Wales; at national level the Patient Register is demonstrated to exceed the Census 2011 population estimate for Wales by 3.1%.

In the absence of Census data, it is conceivable that a SAIL based system could be used to construct a 100% Wales Longitudinal Study. However, as previously discussed, the value of the Longitudinal Studies largely derives from their ability to combine socio-economic data from the Census with linked administrative records. Apart from the Census, the Annual Population Survey is the largest single source of socio-economic data for Wales. However, the inclusion of consent to link questions within the APS to allow this data to be transferred out of ONS for the purpose of data

linking would appear to be highly problematic in the context of how this survey is conducted. Given the level of WG investment in this data set in terms of boosting the sample size for Wales, consideration should be given to exploring whether the terms and conditions under which this data is collected could be changed in a way such that greater value added could be extracted from this data by the Welsh Government. Attention should also be given to the development of a set of harmonised core questions that would be required to appear in all Welsh Government surveys.

Chapter 5: An Overview and Application of Statistical Matching

5.1 Introduction[76]

An important innovation of the proposed Welsh Longitudinal Study is to maximise the use of retrospective, anonymised survey data through statistical matching. The benefits of linking administrative records held for the same individual in different data sets (via uniquely identifying personal information such as NHS registration numbers), such as the linking activity that underpins the existing three UK Longitudinal Studies, are well established. Whilst such 'direct linking' would be an integral component of a Welsh Longitudinal Study, the development of a WLS also aims to go a step further and use statistical matching techniques to match survey records for anonymous individuals to the most similar individual (based on a variety of personal and socio-economic characteristics) within the WLS. For example, among those respondents who give consent for data linking, rich information on the health of the Welsh population collected via the Welsh Health Survey can be linked directly to the spine of the WLS. However, the WHS is a relatively small survey and therefore a majority of people within the WLS will not have responded to the WHS. For this group (and for those who participated in the WHS but did not give their permission for data linking), it is proposed that the 'gaps' can be filled by finding the closest match from the WHS to create simulated data for that individual. Over time, simulated data for individuals within the WLS can be replaced by real data among those who respond to the WHS.

This chapter provides an overview of statistical matching and provides an example of statistical matching to illustrate some of the issues that need to be considered when undertaking such matching exercises. Section 5.2 firstly provides an overview of statistical matching and how statistical matching is operationalised in practice via the technique of Propensity Score Matching. Section 5.3 outlines the practical issues that need to be considered when undertaking such matching exercises; most significantly the availability of a good selection of consistently defined variables within both of the data sets that are to be matched. Section 5.4 considers some of the methodological issues pertinent to statistical matching. Finally, Section 5.5 provides an example of statistical matching in practice. By way of illustration, data on poverty collected for households via the Family Resources Survey (FRS) and held on the Households Below Average Income data set (HBAI) is statistically matched on to the Labour Force Survey. This example serves to highlight the

[76] This chapter incorporates data from the Labour Force Survey and the Households Below Average Income data set which are produced by the ONS and is accesses via end user license from the UK Data Archive, University of Essex, Colchester. None of these organisations bears any responsibility for the analysis or interpretation undertaken here.

practical issues that are encountered, the sensitivity of results to different assumptions made with respect to matching and how the technique can be used to develop new combinations of variables and thereby opportunities for analysis.

5.2 An Overview of Statistical Matching

The purpose of statistical matching is to append new 'synthetic' information to a data set to create new combinations of data. The following definition is provided by Kum and Masterson (2008)[77].

'Statistical matching is a technique used to link records in two separate data sets in cases when exact matching of individual records (record linkage) is not possible due to confidentiality restrictions on the data available. Statistical matching uses variables common to both data sets to identify similar records that can be linked in order to generate a new synthetic data set that allows more flexible analysis than would be possible with the two discrete data sets' (Kum and Masterson, 2008, p1).

The two data sets involved in the matching process can be regarded as constituting a 'donor' data set and a 'recipient' data set. Subject to information constraints, statistical matching can be operationalised within a variety of settings. A commonly used application of statistical matching is in the field of Counterfactual Impact Evaluation, where statistical matching techniques are used to develop control groups for individuals who have participated in different types of intervention (e.g. medical, labour market etc). Any appraisal of the impact of an intervention requires an account of what would have happened to participants if they had not received the intervention (known as the counterfactual). A worthwhile counterfactual implicitly defines a control group or sample whose experiences accurately reflect the hypothetical, unobserved outcomes for the treatment group. Ideally individuals would be allocated to the control and treatment groups at random before participation in an intervention. Outside of medical research, this ideal is rarely achieved in practice and statistical matching techniques have been developed to provide methods for defining control groups and evaluating treatments in the absence of an initial ideal experimental allocation (see Rosenbaum, 2002[78]; Caliendo and Kopeinig, 2008[79] for introductions to statistical matching). The idea behind statistical matching in this context is simply to select a group of non-participants in a way that makes them resemble the participants in every characteristic but the fact of receiving the intervention. If this is done accurately then the outcome observed for the matched group approximates the counterfactual (i.e. what the participants would have done in the absence of the

[77] Hyunsub Kum & Thomas Masterson, 2008. "Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Wellbeing," Economics Working Paper Archive wp_535, Levy Economics Institute.
[78] Paul R Rosenbaum, 2002 "Observational Studies" Spring Series in Statistics, New York.
[79] Marco Caliendo and Sabine Kopeinig (2008) "Some Practical Guidance for the Implementation of Propensity Score Matching", Journal of Economic Surveys, 22(1) pp32-71.

intervention), and the effect of the intervention is straightforwardly estimated as the difference between the average outcomes of the two groups.

Within Wales, Counterfactual Impact Evaluation techniques have been used in examining the employment outcomes associated within participation in ESF (European Structural Funds) funded training programmes (see Davies et al 2012[80]). As part of their monitoring and evaluation activities, the Welsh European Funding Office commission surveys of leavers from ESF funded training programmes. These surveys collect detailed information on the characteristics, experiences and subsequent labour market outcomes of ESF participants. However, other than the subjective assessments of respondents as to the impact of ESF, such surveys of participants are not able to demonstrate what would have happened to ESF participants in the absence of ESF. To provide a counterfactual, the respondents to the ESF survey are statistically matched to respondents to the Labour Force Survey who share similar characteristics to participants in ESF. Both surveys provide information on transitions in economic activity measured over a period of 12 months. It is therefore possible to compare the rate with which previously unemployed or economically inactive ESF participants enter employment following the completion of their training programme with the rates observed among comparable groups of unemployed and economically inactive people within the wider population. It is also noted that in the context of evaluating ESF, statistical matching was undertaken on specific sub-groups of the population (i.e. the unemployed or economically inactive) as opposed to the entire sample of LFS respondents. Such restrictions in themselves enhance the likelihood that 'like with like' comparisons are being made.

In the case of Counterfactual Impact Evaluation, the purpose of statistical matching is to merge new variables that approximate the counterfactual for a given individual. Within the context of evaluation ESF, this process is implemented by appending records of respondents to the LFS who are comparable to ESF participants to survey responses collected via the ESF Leavers Survey. Employment transition data from the LFS forms essentially provides a new 'hypothetical employment transition' variable that is merged on to the data provided by ESF participants. For each ESF participant, this new variable is populated with data from the LFS respondent that is identified as the closest match. Subject to assumptions regarding the effectiveness of statistical matching, differences in these rates of transition in to paid employment can be interpreted as representing the impact of these programmes on participation in employment.

In this case, the `donor' data set (LFS) was large in comparison to the 'recipient' data set (ESF). Alternatively, statistical matching can also be used to append information from a relatively small scale survey that contains richer information on a particular topic area to a larger survey that contains more limited information related to a larger sample of respondents. In this case, the

[80] Davies R., Makepeace, G., Munday, M., Williams G. and Winterbotham M. (2012), 2010 ESF Leavers Survey, Welsh European Funding Office: http://wales.gov.uk/docs/wefo/report/1206112010esfleaverssurvey2010en.pdf

‘donor’ data set would be small in comparison to the ‘recipient’ data set. To provide enhanced data for all individuals in the ‘recipient’ data set therefore implies that information from the same individual within the ‘donor’ data set is allocated to multiple people within the ‘recipient’ data set. This is achieved by allowing matching ‘with replacement’, where an individual from the donor data set who has been ‘matched’ is allowed to be returned to the pool of donor data from which further matches can be drawn. It is therefore important to note that statistical matching does not serve as a replacement for the grossing factors that may be contained within the smaller ‘donor’ data set. The repeated allocation of data from the same individual within the ‘donor’ data set does not necessarily facilitate analysis for more detailed sub-groups of the population than would be possible based upon the original information contained within the ‘donor’ data set alone. The value of statistical matching is the ability to conduct analysis of new combinations of data items in situations where direct linkage is not feasible.

5.3 Operationalising Statistical Matching: Propensity Score Matching

Introducing Propensity Score Matching

Matching is sometimes referred to as 'selection on observables'. Suppose that in two data sets that were selected to act as a 'donor' and 'recipient' data set, it is assumed that all that was known about respondents in both data sets was their gender and whether or not they had a qualification. Statistical matching would involve finding a comparable person in the donor data set for each person in the recipient data set. This simplistic example is referred to as one to one matching. In this example, a male with a qualification in the donor data set would be matched with a male with a qualification in the recipient data set. This example involves exact matching on characteristics (matching on covariates). One practical problem with matching is the 'curse of dimensionality'. This occurs when one or more of the attributes takes many different values and there are several different attributes. The previous example matched on 4 values or cells (2 genders $\times$ 2 qualification levels) and is unlikely to produce reliable results. It would be expected that the availability of a more comprehensive selection of demographic variables coded to a finer degree of detail would produce more accurate matches. However, if we moved to 45 values for age (e.g. a continuous age measure that covered a sample of the population aged 20-64 years), six levels for qualification (e.g. NVQ equivalents plus a category for those with no qualifications) and three locations for place of birth as well as gender, then the match would be on 1350 cells. Such levels of detail would make it much harder to achieve substantial numbers of good matches. Such a problem is likely to be severe because social surveys generally contain many different measurable attributes relating to respondents that we would wish to include as characteristics to be accounted for within any statistical matching exercise.

Propensity score matching (PSM) resolves the 'curse of dimensionality' by estimating the probability of a respondent appearing in the donor data set and creating a matched case in the recipient data set by matching individuals from the donor data set who have similar propensity scores (Rosenbaum and Rubin, 1983[81]). The propensity score can be derived from a statistical model that estimates the probability of being in the donor data set based upon their observable characteristics. In terms of implementing this, data from the donor data set is appended to data from the recipient data set for those data items that appear in both surveys. Individuals in the combined data set are distinguished in terms of whether their response came from the donor data set or whether it came from the recipient data set on the basis of simple dichotomous 0/1 variable.

[81] Paul R. Rosenbaum and Donald B. Rubin (1983). The Central Role of Propensity Score in Observational Studies for Causal Effects, Biometrika, 70(1), pp41-55

Analysis is then undertaken to estimate what characteristics are associated with the probability of an individual in the combined data set having come from the original donor data set through the use of multivariate statistical analysis, such as a probit or logit regression (i.e. techniques that are suitable for a dichotomous dependent variable). This process allows the identification of the characteristics that are associated with people being more or less likely to be included in the donor data set (as measured by their propensity score). Respondents from the 'donor' and 'recipient' data sets can then be matched on the basis of their propensity score; their estimated probability of being in the 'donor' data set given their observable characteristics.

Conditional Independence Assumption

The key assumption made in matching models is the Conditional Independence Assumption (CIA) also known variously as ignorability and unconfoundedness. Respondents in the donor and recipient data sets may differ because they have different characteristics. For example, different files may be expected to display different characteristics if, for example, the data sets were derived from social surveys that varied in terms of their sampling designs, such as the oversampling of certain population sub-groups such as ethnic minorities. Some of these characteristics (e.g. gender or age) are observable and can be used as control variables to adjust for differences between the groups. Other characteristics may, however, be unobservable. For example, voluntary surveys with low response rates may indicate the presence of response bias insofar that responses may not have been achieved with sub-groups of the population that are relatively hard to reach. Whilst such differences may largely be able to be accounted for by matching on the basis of observable characteristics, it may remain the case that respondents differ in terms of their 'unobservable' willingness to respond to such surveys. However, PSM has to assume that these unobservables do not have a systematic effect on the characteristics of respondents from the 2 surveys. More formally, records in both files are assumed to be drawn randomly and independently of each other from the sample population. The CIA is a statement of conditions under which the effects of the unobservables can be ignored.

Common Support

An attempt is made to match a respondent from a donor data set to a respondent in a recipient data set on the basis of their probability of appearing within the donor data set, where this probability is calculated on the basis of observed individual characteristics. However, it may not be possible to successfully match all cases within the recipient data set. The existence of a substantial overlap between the characteristics of recipients and donors (or the treated and non-treated groups in the context of observational studies), referred to as 'Common Support', is another requirement for the applicability of this method. The common support is the domain over which the groups in the 'donor' and 'recipient' files are directly comparable. In simple terms, it is the set of

individuals in the data sets who share similar values of the control variables used in the matching process and who could reasonably be expected to appear in either data file. In matching surveys that are expected to be representative of the wider population, it should be possible to match individuals who are similar in both sources. Lack of common support is more likely to arise where PSM techniques are being used to construct control groups for the purposes of Counterfactual Impact Assessments where those who are in receipt of a 'treatment' are unrepresentative of the wider population.

5.4 Practical Issues Surrounding Propensity Score Matching

The quality of statistical matching depends upon the availability of a *comprehensive selection of common and consistently coded variables.* It is therefore important to note that the information required to facilitate accurate statistical matching will vary depending upon the subject matter being considered. A 'comprehensive set of variables' refers to the selection of measures that underlying theory and previous research would suggest should be included within any analytical model that seeks to predict the value of an outcome measure. For example, household structure is an important determinant of household income and, in turn, whether or not a household is observed to be living in poverty. The presence or absence of children within a household is particularly important in the derivation of 'equivalised household income' within the Households Below Average Income (HBAI) data set; the official source of data on the incidence of poverty in the UK. Using the HBAI as a 'donor' data set to effectively impute measures of equivalised household income on to a 'recipient' data set would require the recipient data set to also contain detailed information on household structure. All other things being equal, a statistical matching exercise would wish to avoid merging the poverty status of a household consisting of a couple with no children from one data set on to a record relating to a lone parent with three children in another data set. This is only likely to be achieved if detailed information on household structure is available for both data sets. The availability of a comprehensive selection of consistently coded variables has implications for the use of statistical matching within the context of the proposed WLS. Even with the inclusion of Census data, the population spine will contain a relatively limited range of demographic and socio-economic variables. The ability to meaningfully match survey data directly on to the population spine may therefore be limited.

To implement statistical matching, data from the 'donor' and 'recipient' data sets need to be pooled. This can only be achieved for variables that are common to both data sets and which are consistently coded. Variables that cover the same subject matter but which are not coded to the same classification schema will not be able to be used unless one or both of the variables from the data sets to be matched can be re-coded in a way that makes them consistent. The chances of being able to construct such consistent variables will be increased if the classification of data within

the ‘donor’ and ‘recipient’ data sets conforms to nationally agreed standards (e.g. categorical data for ethnicity, social class, occupation). Even if information is coded to different levels of detail (e.g. those for social class available through the National Statistics Socio-Economic Classification: NS-SEC), the utilisation of coding frames that comply with national standards should mean that the level of detail used in one data set can be ‘collapsed’ to be consistent with that used in the other data set.

Further complications may arise if there are significant differences in the way that data is collected. For example, the previous chapter identified differences in the way that data on social class is measured within the Welsh Health Survey (highest income householder) compared to the National Survey for Wales (survey respondent). Merging information on socio-economic status from these 2 sources is therefore problematic, to the extent that social class may not be able to be used for the purpose of statistical matching. Such issues mean that the methodologies underpinning different sources of data need to be examined to consider whether even identically coded variables are consistently measuring the same thing.

5.5 Methodological Issues Surrounding Propensity Score Matching

Conceptually, the simplest type of Propensity Score Matching (PSM) is nearest neighbour matching. The nearest neighbour of a person in the recipient sample is the person in the donor sample that is the smallest distance away in terms of the propensity score.[82] This criterion may result in poor matches so a calliper is often specified which defines a maximum acceptable difference between the two propensity scores and therefore rejects matches that are regarded as too dissimilar. Where a match is made, the statistical procedure allocates a unique id number relating to the case within the donor data set that is acting as the nearest neighbour and the value of the variable being that is being donated to the recipient.

A common practical problem is what to do when there are relatively few people in the donor sample compared to the recipient sample. Matching without replacement makes the closest match between the recipient and donor data set and removes the corresponding observation from the data set from the list available for matching. Matching with replacement allows each observation in the donor data set to be potentially matched to more than one observation in the recipient data set. After each match is made, the observation from the donor data set is returned to the pool available for matching. Radius matching is a further refinement. Here, each donor observation is allocated the average value of a variable derived from all the observations in the recipient data set within the distance specified by the calliper (in contrast to nearest neighbour matching the id numbers of the donors are not allocated). Finally, it may be appropriate to segment the data files and perform

[82] The measure of distance is the absolute value of the difference in propensity scores. Other measures of distance are possible.

statistical matching separately for different sub-groups of the population. Matching on the basis of a propensity score infers matching on the basis of the 'combined' characteristics of individuals and does not mean that like-for-like matches (as demonstrated in the description of one-to-one matching earlier in this chapter) are always made. For example, a record for a female in the 'donor' data set could quite conceivably be matched to the record for a male in the 'recipient' data set if the net effect of all the matching characteristics produced the closest propensity score (i.e. the differences associated with gender were offset by other observable characteristics). There is no objective 'test' of the correct method to be used and judgements are required regarding the characteristics of the data sets being matched and the purpose of the research. In practice, the approach generally taken by researchers is to implement a variety of techniques to consider whether the results derived from statistical matching are sensitive to the application of different techniques, the size of callipers, assumptions regarding replacement and the segmentation of data files.

5.6 Matching Poverty and Labour Market Data

This section provides an example of statistical matching in practice. By way of illustration, data on poverty collected for households via the Family Resources Survey (FRS) and held on the Households Below Average Income data set (HBAI) is statistically matched on to data from the household files of the Labour Force Survey. This example serves to highlight the practical issues that are encountered, the sensitivity of results to different assumptions made with respect to matching and how the technique can be used to develop new combinations of variables and thereby opportunities for analysis. At the outset, it is acknowledged that this exercise is not designed to provide a 'gold standard' exemplar of statistical matching in the field of poverty analysis. For example, the matching exercise uses standard, anonymised 'End User Licence' versions of the HBAI and LFS data sets. More disclosive versions of these data sets are available and it is conceivable that the additional detail that is available from these data sets could have enhanced the accuracy of the matching exercise. However, this is not important in the context of the present exercise where the purpose is to provide a worked example of statistical matching in relation to the analysis of poverty. Both the HBAI and LFS data sets include a range of demographic and socio-economic variables that are typical of those included in government surveys and therefore provides a representative illustration of statistical matching using two 'off the shelf' sources of data.

Estimates of Poverty from HBAI

Households Below Average Income (HBAI) data are derived from the Family Resources Survey (FRS) and are regarded as the key dataset for the analysis of poverty within the UK. The FRS achieves full interviews with approximately 23,000 respondents from households in Great Britain

and a further 2,000 households in Northern Ireland. The size of the FRS sample is relatively small, which is problematic in terms of undertaking an analysis of poverty among population sub-groups in Wales. In accordance with HBAI publication standards for the presentation of regional data, the analysis of poverty in Wales that follows is based on 3 years worth of HBAI data, covering the period 2008/9 to 20010/11. All income values have been uprated to 2010/11 levels using the HBAI deflator series[83]. It should also be noted that estimates are derived from a non-disclosive version of the HBAI data that is made available for research use. For these reasons, the results may not exactly match official estimates presented in HBAI publications[84].

The basic unit of analysis within the HBAI data set is the benefit unit. A family, or benefit unit, is a single adult or a couple, together with any dependent children. An adult living in the same household as his or her parents, for example, is defined as a separate benefit unit from the parents. The HBAI uses household disposable incomes, after adjusting for the household size and composition, as a proxy for material living standards. All individuals in the household are assumed to benefit equally from the combined income of the household and are therefore each allocated the same equivalised household income; i.e. the household income is adjusted according to the composition of the household. This enables the total equivalised net weekly income of the household to be used as a proxy for the standard of living of each household member making it easier to compare household incomes in relation to household needs.

The primary purpose of the HBAI data set is to examine how variations in equivalised income between different population sub-groups translate into the proportion of people who live in poverty. Within publications based upon the HBAI, figures are presented on the number of people living in households that have income below certain thresholds of median income, with results being typically presented for less than 50%, less than 60% and less than 70% of median income. Of these measures, the principal marker of low income is generally regarded as households with less than 60 per cent of median income and so this is the definition that is adopted in the analysis of poverty that follows. Official estimates of the number of people living beneath HBAI income thresholds presented both before and after housing costs are taken into account. The treatment of housing costs is particularly important in the context of regional comparisons, where regional differences in income partly reflect differences in the costs of living (e.g. additional allowances received by those working in London and the South East, sometimes referred to as 'London

[83] See document 'HBAI datasets – Guidance for the production and checking of analysis' that accompanies UK Data Archive Study Number 5828: Households Below Average Income, 1994/5-20010/11.

[84] The research version of the HBAI data suppresses the responses provided by some respondents in order to maintain anonymity.

Weighting'). The analysis that follows examines the incidence of poverty in Wales after housing costs have been taken in to account.

Table 5.1 presents estimates of the rate of poverty in Wales among different population sub-groups. To allow information collected at the level of the benefit unit from the HBAI to be merged on to household level data contained within the Labour Force Survey, the analysis is restricted to single benefit unit households in the HBAI. Overall, it is observed that approximately one in five single benefit unit households in Wales are estimated to be living in poverty after housing costs are taken in to account. At the outset it should be noted that the choice of these dimensions, and the derivation of population sub-groups, was driven entirely by the availability of demographic and socio-economic data within the HBAI and LFS data sets and the ability to transform these variables into consistently coded categories. Poverty is highest amongst the young, where approximately 44% of single benefit unit households that are headed by somebody aged 16-24 years are estimated to be in poverty.

Table 5.1: Percentage of Households in Poverty (2008/9-2010/11)

	After Housing Costs (%)	Before Housing Costs (%)
Age band		
16 to 24 yrs	44.1	23.4
25 to 34 yrs	24.5	13.0
35 to 44 yrs	22.8	14.1
45 to 54 yrs	21.7	15.3
55 to 64 yrs	20.7	17.8
65 to 74 yrs	13.6	14.3
75 to 79 yrs	14.7	16.1
80+ yrs	17.3	19.4
Family type		
Male pensioner	13.0	11.7
Female pensioner	15.9	18.3
Pensioner couple	14.8	15.1
Couple with 1 child	17.7	11.4
Couple with 2 children	17.9	12.1
Couple with 3 children	29.4	21.5
Couple no children	13.1	10.2
Lone parent with 1 child	41.5	18.4
Lone parent with 2 children	40.4	18.2
Lone parent with 3 children	43.9	24.7
Single male	32.9	22.8
Single female	30.6	19.2
Economic Status		
Self employed	24.8	20.0
All working full time	5.7	2.6
One full, one part time	6.6	3.6
One full, one non-working	20.8	13.1
One or both part time	26.8	18.7
Workless pensioners	17.4	18.3
Workless, unemployed	90.4	66.7
Workless, inactive	61.3	35.8
Housing Tenure		
Owns outright	13.6	18.6
Own with mortgage	11.4	8.5
Rent council	38.3	23.4
Rent local authority	38.2	19.0
Rent other	38.1	15.1
Rent free	19.1	27.0
Total	**20.9**	**15.6**

Note: Sample restricted to single benefit unit households, data are un-weighted

In terms of family type, poverty is highest within households that are made up of single people and lone parents. The presence (or absence) of children within a household is important in determining

whether or not a household is defined as living in poverty. The rate of poverty among lone parents is approximately 10 percentage points higher than that observed among single-person households. Levels of poverty are lowest amongst couples with no children (13%) and pensioner households. In terms of economic status, over 90% of workless households where at least one person is unemployed are estimated to be living in poverty. This is considerably higher than that observed among the economically inactive (61%), reflecting the heterogeneous nature of this group. Poverty is lowest among households where two people are observed to be in work (between 6 and 7%). Finally, the prevalence of poverty varies by tenure. Amongst renters, the incidence of poverty is estimated to be 38%.

Statistical Matching Based on HBAI Data

The accuracy of any statistical matching exercise will be determined by the coverage and content of variables used in the matching process. The validity with which information on age group, family status, economic status and housing tenure can be used to match information on poverty from HBAI data on to the LFS can be demonstrated by the explanatory power of multivariate statistical models that model variations in the incidence of poverty among population sub-groups based upon these characteristics. Based upon 2010 HBAI data for single benefit unit households, Table 5.2 summarises results of three logistic regressions that estimate the probability of a household being in poverty. The specification of explanatory variables included within these models replicate the categories presented in Table 5.1. It can be seen that the statistical model of poverty that includes only variables relating to age group and family type has a very low explanatory power (Pseudo R-Squared of 0.05) compared to the models that also include economic status (0.20) and both economic and housing status (0.22).

To examine the effect of this increase in explanatory power on statistical matching, a matching exercise was undertaken on 2009 and 2010 HBAI data using PSM techniques. The poverty status of single benefit unit households contained the 2010 HBAI data were matched on to records contained within the 2009 data. The recorded poverty status of households in the 2009 HBAI data could then be compared with their matched poverty status as derived from the 2010 data. The technique applied is Nearest Neighbour matching without replacement (calliper=0.0001). Table 5.2 reveals that, overall, the level of consistency between actual and imputed poverty appears to be in the order of 70-75%. However, it is noted that only 20% of households live in poverty and therefore high rates of consistency would therefore have been expected. Among those households who were recorded as being in poverty in 2009, it can be seen that only 1 in 4 are also imputed to be in poverty based upon a statistical matching exercise that controls for age group and family type. The inclusion of additional information on economic status increases the level of consistency to 40%. Nonetheless, it remains the case that based upon a relatively parsimonious statistical model, only a minority of those households recorded as being in poverty are matched to

households who are also recorded as being in poverty. Such levels of inconsistency derived from imputed results are not uncommon. Previous research based on the ONS LS has considered the levels of consistency between ethnicity recorded in the 1991 Census with an imputed ethnicity variable created for the same respondents in the 2001 Census record[85]. Ethnicity was imputed for approximately 3% of respondents to the 2001 Census where ethnicity data was either missing or invalid. Of those from 1991 minority ethnic groups, research revealed that less than half were imputed to the same ethnicity as they used in their 1991 Census response. It must be noted that the purpose of the ethnicity imputation was to produce population estimates and the integrity of the imputed data at the individual level was not the main priority. Nonetheless, the exercise demonstrates the inconsistencies that can emerge with statistical matching.

Table 5.2: Summary of Performance of Logistic Regressions

	Model 1	Model 2	Model 3
Diagnostics			
Pseudo R-Squared	0.053	0.203	0.223
% Correct predictions from statistical matching			
All	69.1%	75.4%	75.5%
Among those in poverty	25.1%	39.3%	38.4%
Model specifications			
Model 1: Controls for age group and family type			
Model 2: Controls for age group, family type and economic status			
Model 3: Controls for age group, family type, economic status and housing tenure			

Table 5.3 presents estimates of the imputed rates of poverty for different population sub-groups based upon each of the three specifications of statistical model. It can be seen that where explanatory variables can be introduced for a particular characteristic (e.g. family status), imputed rates of poverty for particular population sub-groups defined by that characteristic (e.g. lone parents) are very similar to actual rates of poverty. Where explanatory variables for a particular dimension are not included (highlighted by italics in Table 5.3), then relatively large differences are observed to exist between imputed rates and actual rates for that characteristic. For example, Model 1 does not contain explanatory variables related to economic status. As a consequence, estimates of poverty derived from matched observations do not align closely with actual levels of poverty observed among households according to their economic status. Similarly, Model 2 does not contain explanatory variables related to housing tenure. Consistency in imputed and actual rates of poverty across each group of household characteristics is highest when imputed rates are

[85] Platt L, Akinwale B, Simpson L. (2006) Stability and change in ethnic group in England and Wales, Population Trends 121: 35-45 available from: http://www.ons.gov.uk/ons/rel/population-trends-rd/population-trends/no--121--autumn-2005/index.html

derived from statistical matching models that control for the widest range of household characteristics.

Matching Poverty on to the Labour Force Survey

The above analysis of HBAI data has revealed that it is possible to 'donate' information about poverty collected from respondents to the Family Resources Survey from one year to respondents to the FRS during another year via a process of statistical matching. Using the same techniques, it is also possible to impute rates of poverty derived from HBAI data sets to other survey or administrative sources that do not themselves contain information on household income. To illustrate this process, this section demonstrates how statistical matching can be used to donate information related to poverty from the HBAI data set to survey responses collected from the Labour Force Survey.

The LFS is the largest regular household survey conducted in the UK. Interviews are conducted in approximately 60 thousand households each quarter, far more than the 24 thousand interviews achieved annually within the Family Resources Survey. The LFS collects information on personal characteristics, household structure, economic activity, health, education and training and earnings. For those in employment, detailed information is collected on jobs held including occupation, hours worked and contractual status. Whilst information is also collected on earnings, the LFS does not collect detailed information about other sources of household income (such as benefits, financial assets etc) and it is therefore not possible to derive information on poverty directly from this source. The allocation of poverty status to LFS respondents based on statistical matching with HBAI data therefore provides a useful illustration of how the technique can be applied.

Table 5.3: Comparing Imputed* and Actual Poverty Rates

	HBAI 2009	Imputed Poverty Rates		
		Model 1	Model 2	Model 3
Age band				
16 to 24 yrs	44.3%	43.8%	45.4%	42.4%
25 to 34 yrs	24.7%	24.7%	24.2%	23.4%
35 to 44 yrs	22.3%	22.8%	21.9%	21.3%
45 to 54 yrs	21.6%	22.6%	21.8%	20.3%
55 to 64 yrs	20.3%	20.7%	20.1%	20.8%
65 to 74 yrs	13.3%	13.8%	13.8%	14.5%
75 to 79 yrs	14.7%	15.0%	14.4%	13.8%
80+ yrs	16.6%	16.0%	16.1%	16.1%
Family type				
Male pensioner	12.7%	12.6%	12.7%	13.0%
Female pensioner	15.5%	15.3%	15.2%	15.5%
Pensioner couple	14.5%	14.7%	14.6%	15.4%
Couple with 1 child	17.3%	17.1%	16.9%	16.5%
Couple with 2 children	18.1%	18.6%	18.0%	17.2%
Couple with 3 children	28.2%	28.8%	27.4%	24.0%
Couple no children	13.2%	13.2%	12.6%	13.4%
Lone parent with 1 child	40.3%	39.0%	38.6%	37.7%
Lone parent with 2 children	38.0%	41.2%	40.0%	37.5%
Lone parent with 3 children	40.6%	43.9%	44.0%	43.0%
Single male	33.7%	34.5%	33.7%	31.1%
Single female	30.5%	32.3%	29.7%	30.3%
Economic Status				
Self employed	24.4%	*20.1%*	24.3%	22.8%
All working full time	5.7%	*24.0%*	6.2%	7.6%
One full, one part time	6.0%	*17.4%*	6.4%	7.2%
One full, one non-working	20.0%	*16.9%*	19.6%	20.1%
One or both part time	25.7%	*24.7%*	25.9%	25.0%
Workless pensioners	17.1%	*15.5%*	16.9%	16.8%
Workless, unemployed	89.5%	*30.2%*	87.9%	76.7%
Workless, inactive	60.2%	*35.2%*	61.3%	58.5%
Housing Tenure				
Owns outright	13.2%	*17.0%*	*16.2%*	13.2%
Own with mortgage	10.9%	*20.6%*	*13.8%*	11.2%
Part own part rent	15.3%	*21.4%*	*12.3%*	17.2%
Rent council	38.0%	*24.5%*	*33.6%*	37.5%
Rent local authority	37.9%	*26.9%*	*33.5%*	36.4%
Rent other	38.2%	*25.5%*	*28.5%*	36.5%
Rent free	19.4%	*22.7%*	*17.1%*	22.9%
Total	20.6%	20.9%	20.3%	19.9%

*Explanatory variables used within these models are the same as those described in Table 5.2.

In matching LFS respondents with those HBAI respondents who share similar characteristics, the first practical issue to be addressed is the identification of a set of variables that are available in both sources of data and which are classified (or can be recoded to be classified) on a consistent basis. The selection and derivation of variables described in Table 5.1 and Table 5.3 has been made with the purpose of statistical matching in mind. In contrast to the HBAI, the unit of analysis within the quarterly LFS data files is typically the individual. However, twice a year LFS data sets including household level data are produced. Whilst the individual remains the unit of analysis within these data sets, the household files also include a household level serial number which allows information from different individuals within the same household to be manipulated and aggregated to construct data at a household level. To ensure that the data to be matched is contemporaneous with the HBAI data described above, three data files relating to the October-December Quarter of 2008, 2009 and 2010 are combined. However, no stratification of the data is implemented and it is therefore possible that a HBAI respondent for 2010 could be matched to an LFS respondent for 2008. For consistency with the HBAI, the sample is restricted to households with one family unit. HBAI data was restricted to single benefit unit households and it therefore must be acknowledged that family units are not necessarily synonymous with benefit units. A family unit may contain more than a one benefit unit, for example where a non-dependent child lives within the parental home.

The HBAI and LFS samples are described in Annex 3. The analysis is restricted to individuals living in Wales. Due to the respective size of the two surveys, the sample size available for Wales from the LFS (6,644 households) is more than twice that derived from the HBAI (2,968 households) for the same three year period. No attempt is made within this analysis to take account of the rotational design of the LFS sample and it is therefore acknowledged that pooling LFS data over a period of 3 years will result in some households appearing twice in the combined data set. In terms of their characteristics, the HBAI and LFS samples appear to be broadly comparable. The one exception is the higher proportion of households in the HBAI sample who are classified as workless pensioners compared to the LFS. The reason for this discrepancy is not clear, although differences in the definitions of benefit units and family units may be contributing to the lower proportion of pensioners generally within the LFS sample.

The matching technique applied to the pooled HBAI/LFS data is Nearest Neighbour matching. Two sets of models are estimated to reflect ‘with replacement’ and ‘without replacement’ matching. ‘With replacement’ allows the same HBAI case to be matched to multiple cases within the LFS data; given the relatively large size of the LFS sample, a necessary condition if matches are to be achieved with all LFS respondents. The matching model utilises the full set of explanatory variables that are available from both sources of data (age, family status, economic status and tenure). The effects of changing the size of calliper are also examined. As noted above, the

calliper specifies the range of predicted probabilities within which acceptable matches can be made. Three models are estimated that employ a) no calliper; (b) a calliper of 0.0001 and (c) a calliper of 0.00001. Model (c) therefore represents the most stringent conditions under which a statistical match would be made. The effects of utilising different callipers on both imputed poverty rates and the size of the matched sample are shown in Table 5.4.

Several key findings emerge:

- Rates of imputed poverty within the LFS are generally lower than actual rates of poverty derived from the HBAI. Differences in overall rates of poverty cannot be fully explained by differences in the composition of the two samples as imputed rates of poverty observed for particular population sub-groups are also observed to be lower within the LFS, although those within group differences are smaller.

- It is generally the case that imputed rates of poverty more closely reflect HBAI estimates when statistical matching does not allow for replacement. Further examination of the matched records reveal that 7 HBAI cases were each being used as a statistical match in more than 100 cases (Calliper 0.0001). In total, these 7 cases accounted for approximately 15% of all statistical matches made. Further examination revealed that 6 of these HBAI households were pensioner households, none of whom were recorded as being in poverty. The example highlights how the overall imputed rate of poverty will therefore be sensitive to the poverty status of just a handful of households within the HBAI data set.

- Across a majority of population sub-groups, the implementation of a smaller calliper generally results in imputed rates of poverty derived for the LFS being closer to actual poverty rates estimated from the HBAI data. However, not allowing for replacement and the utilisation of a smaller calliper means that the sample size of LFS respondents for whom a suitable match can be derived is considerably lower.

Table 5.4: Imputed Poverty Rates for the LFS:

		With Replacement			No Replacement		
	HBAI 2008/10	No Calliper	Calliper 0.0001	Caliper 0.00001	No Calliper	Calliper 0.0001	Caliper 0.00001
Age band							
16 to 24 yrs	47.3%	26.4%	30.5%	30.8%	20.9%	40.4%	37.1%
25 to 34 yrs	24.9%	20.8%	10.3%	7.2%	24.6%	15.9%	15.8%
35 to 44 yrs	21.6%	19.2%	17.9%	21.3%	19.5%	19.7%	19.7%
45 to 54 yrs	25.0%	16.3%	19.0%	21.6%	24.6%	21.0%	20.1%
55 to 64 yrs	21.8%	23.6%	13.5%	19.9%	31.3%	19.3%	19.2%
65 to 74 yrs	13.6%	13.3%	15.1%	3.4%	17.8%	14.3%	12.9%
75 to 79 yrs	9.2%	5.7%	11.0%	3.9%	17.9%	12.0%	9.1%
80+ yrs	17.9%	11.7%	8.9%	37.9%	17.5%	18.4%	22.4%
Family type							
Male pensioner	12.9%	11.3%	3.7%	32.3%	24.9%	14.4%	11.1%
Female pensioner	14.1%	14.0%	8.0%	2.8%	15.2%	13.5%	16.7%
Pensioner couple	15.7%	13.5%	14.6%	15.8%	20.1%	17.0%	15.7%
Couple with 1 child	18.1%	21.4%	11.8%	4.0%	21.7%	12.8%	11.7%
Couple with 2 children	19.4%	9.2%	16.3%	17.1%	18.4%	13.8%	13.7%
Couple with 3 children	31.7%	35.3%	26.3%	38.7%	24.4%	27.7%	28.7%
Couple no children	11.8%	12.2%	12.7%	16.2%	18.0%	11.1%	10.1%
Lone parent with 1 child	47.7%	20.5%	22.1%	24.0%	35.7%	41.4%	42.7%
Lone parent with 2 children	39.1%	22.7%	20.4%	21.5%	35.1%	34.6%	29.0%
Lone parent with 3 children	56.9%	43.3%	28.7%	49.7%	31.8%	44.0%	52.8%
Single male	37.8%	28.0%	17.3%	29.4%	31.5%	28.2%	28.1%
Single female	34.9%	29.5%	29.4%	28.9%	40.1%	29.5%	29.9%
Economic Status							
Self employed	23.3%	23.1%	33.4%	30.6%	16.3%	21.5%	21.3%
All working full time	5.9%	9.0%	6.7%	5.7%	16.8%	6.5%	5.5%
One full, one part time	8.1%	14.1%	9.5%	17.7%	40.4%	8.9%	10.0%
One full, one non-working	19.2%	39.2%	33.2%	3.2%	14.3%	18.6%	17.2%
One or both part time	31.9%	20.4%	15.3%	19.6%	19.4%	26.0%	27.6%
Workless pensioners	16.8%	12.2%	6.8%	19.2%	19.2%	17.2%	17.7%
Workless, unemployed	92.0%	76.6%	94.5%	94.5%	67.1%	93.0%	93.0%
Workless, inactive	63.3%	66.1%	62.9%	65.9%	46.1%	61.9%	61.8%
Housing Tenure							
Owns outright	14.8%	9.8%	9.6%	13.8%	18.7%	14.6%	14.6%
Own with mortgage	10.7%	14.3%	11.2%	15.8%	19.8%	9.6%	9.5%
Rent council	38.8%	26.8%	23.4%	23.0%	25.1%	31.0%	30.8%
Rent local authority	45.6%	43.5%	43.3%	28.9%	27.1%	38.8%	37.2%
Rent other	47.7%	31.3%	28.2%	25.9%	30.4%	40.2%	42.0%
Rent free	14.0%	20.6%	7.6%	20.7%	32.9%	12.9%	12.2%
Total	21.5%	17.6%	15.0%	17.3%	21.1%	17.8%	17.7%
Sample	2,968	6,643	6,044	5,810	2,968	2,605	2,575

Analysis of Poverty by Social Class from the LFS

A potential benefit of statistical matching is that the generation of synthetic data enables the analysis of data from 2 separate sources to be combined without any direct linking of the data set having been required. A potential application of this is demonstrated in Table 5.5. The HBAI EUL data set does not contain any information on the social class of the household. The LFS, however, does contain information on social class. Conceptually, social class aims to differentiate positions within the labour market as defined by the typical employment relations held by people in similar types of jobs. Within a market economy, social relationships are expressed through labour market relationships and employment contracts. The imputation of poverty status on to cases within the LFS therefore provides the opportunity to examine whether social class gradients exist in relation to poverty.

Table 5.5: Poverty by Social Class

	With Replacement			**No Replacement**		
	No Calliper	**Calliper 0.0001**	**Caliper 0.00001**	**No Calliper**	**Calliper 0.0001**	**Caliper 0.00001**
Eight Class Version						
Higher managerial and professional	13.5%	11.8%	14.2%	18.7%	11.2%	9.4%
Lower managerial and professional	11.9%	10.5%	11.0%	19.8%	8.8%	9.8%
Intermediate occupations	15.1%	10.4%	13.4%	23.2%	15.4%	13.1%
Small employers and own account workers	22.3%	30.5%	26.2%	16.9%	23.2%	18.5%
Lower supervisory and technical	13.8%	9.2%	9.3%	18.6%	9.3%	7.9%
Semi-routine occupations	19.1%	16.1%	14.4%	15.0%	14.6%	18.3%
Routine occupations	16.8%	13.5%	10.8%	16.8%	17.8%	16.5%
Never worked, unemployed, and nec	21.9%	16.2%	24.2%	22.9%	22.7%	23.2%
Three Class Version						
Higher managerial, administrative and professional occupations	12.5%	11.0%	12.3%	19.4%	9.7%	9.6%
Intermediate occupations	19.9%	23.6%	21.8%	19.1%	20.3%	16.5%
Routine and manual occupations	16.6%	12.9%	11.5%	16.9%	13.6%	13.9%
Never worked, unemployed, and nec	21.9%	16.2%	24.2%	22.9%	22.7%	23.2%
Overall	17.6%	15.0%	17.3%	21.1%	17.8%	17.7%
Sample	6643	6044	5810	2968	2605	2575

A clear limitation of the matching exercise is that information about the nature of employment, such as hours worked, contractual status or earnings, is not included in the selection of variables used for statistical matching. This is because these variables are not included within the HBAI data. Each of these would be expected to be correlated with social class and would have been expected to enhance the accuracy of the matching exercise. In the absence of these measures, the

estimated relationship between social class and poverty is being underpinned by the extent to which age, family status, economic status and tenure are also related to social class. Given these limitations, it may be too much to expect a clear social class gradient to emerge across the full 8 class version of the National Statistics Socio-Economic Classification (NS-SEC). The analysis is therefore also repeated for the three class version of NS-SEC.

Statistical matching models that allow for replacement do not reveal the presence of any differences in the rate of poverty between households headed by those in Managerial and Professional Occupations and those households headed by people in Routine and Manual occupations. Rates of poverty are highest among those in Intermediate Occupations and those who are unemployed or who have never worked. This obviously calls into question the accuracy of statistical matching with respect to social class, given the set of matching variables that are available. However, matching models that allow for replacement appear to produce more valid results, particularly where the statistical matching model employs a relatively restrictive calliper. Results derived from these models suggest that rates of poverty are lowest among households headed by those in Managerial and Professional Occupations and highest among those households that are headed by the unemployed or those who have never worked; findings that would appear to be more intuitive,

5.7 Concluding Comments

An important innovation of the proposed Welsh Longitudinal Study is to maximise the use of retrospective, anonymised survey data through statistical matching. The quality of statistical matching depends upon the availability of a *comprehensive selection of common and consistently coded variables* in both data sets that are to be matched. In terms of the application of statistical matching within the context of the WLS, an essential requirement relates to the availability of a population spine that contains sufficient detail on the socio-economic characteristics of individuals to facilitate statistical matching to other sources. The exemplar analysis of imputing poverty status on to the Labour Force Survey provides an indication of the level of detail that would need to be embodied within the population spine to facilitate sensible statistical matching. Whilst further refinements to that analysis could be made (e.g. matching within strata), the implications of the exercise are clear: meaningful statistical matching could not be undertaken if the population spine of the WLS is based upon the NHS Administrative Register which only contains information on age, gender and geographical location. Whilst the direct linking of additional administrative sources of data to the NHSAR will provide further information on the population, such information can only contribute to statistical matching if such data is also available in a donor data set. Inconsistencies in how this data is collected and recorded would also have to be considered. Even with the inclusion of Census data, the population spine would only contain a relatively limited range

of demographic and socio-economic variables. The suitability of these variables for statistical matching would need to be considered on a case by case basis.

It is noted that regression modelling provides an alternative approach to imputation. In the context of the present example, a regression model that predicted whether or not a household was in poverty would have been estimated on the HBAI data. The parameters derived from this model could then have been used within the LFS data to provide a prediction as to whether or not a LFS household was likely to be in poverty. Although the approach differs, the requirements to produce sensible estimates remain unchanged; i.e. a comprehensive and consistently coded selection of variables available within both data sets.

The ability to meaningfully match survey data directly on to the population spine may therefore be limited. The performance of statistical matching is therefore likely to be better when survey data can be matched directly to other sources of survey data, as such sources are likely to contain a richer range of variables that can be incorporated in to the matching process. Instead of matching surveys to the available population spine on a case by case basis, a more effective approach may be to statistically match survey data as an exercise which takes place separately to the population. This matching could possibly augmented by any additional administrative data that is available for the two groups of survey respondents if they have provided consent for data linkage, although methods would need to be developed with respect to the matching of non-consenting groups. Such an exercise would then create a composite matched file (e.g. a file containing all statistically matched survey data), the cases in which can be directly linked to the population spine if consent to link has been achieved in one of the surveys. However, such an approach only attempts to enhance existing survey response and does not attempt to 'fill in the gaps' of the population spine for those people who have not responded to a particular survey.

Chapter 6: Conclusions and Recommendations

This report has aimed to establish the key issues that would need to be addressed to construct a Welsh Longitudinal Study that incorporates Census data against which other data sources can be linked or statistically matched. Compared to the other devolved nations of the UK, Wales does not have its own Census-based Longitudinal Study. Instead, the Wales population is included in the ONS England and Wales Longitudinal Study (LS). The ability of both Scotland and Northern Ireland to construct longitudinal studies that incorporate Census data relies on the fact that responsibility for conducting and administering the Census is devolved. Whilst ONS is responsible for undertaking the Census in England and Wales, NRS and NISRA are the custodians of their Census data and have therefore been in a position to develop longitudinal studies in a way that has not been possible in Wales. The main area of research that is conducted using these Longitudinal Studies relates to public health. Given the devolution of powers to the Welsh Government in the field of health, the inability to develop a longitudinal database in Wales to contribute to the evidence base would appear to be inequitable.

The key issue that firstly needs to be addressed is whether the Welsh Government requires a *Census* based Longitudinal Study. The ONS Longitudinal Study was developed to address a particular statistical requirement; to provide better data on occupational mortality and fertility. The initial emphasis of the ONS LS was therefore upon the development and production of aggregate 'Official Statistics' Within both Scotland and Northern Ireland, the impetus to establish Census based longitudinal studies came from their respective academic communities. The impetus behind the SLS and NILS was the establishment of a research resource that could be used to answer specific research questions. Irrespective of the reasons surrounding their establishment, each of the UK Longitudinal Studies has developed into an important research resource. The growth of communities of researchers that use these facilities and the expertise that they develop and share is of mutual benefit to both the government and academic sectors.

The decision as to whether the inclusion of Census data is an integral requirement to the creation of a Wales Longitudinal Study would be likely to shape how and where such a study would be developed. As custodians of Census data for Wales, the development of a study that incorporated Census data would need to be undertaken in collaboration with the Office for National Statistics. In essence, the WLS would comprise of a 'boost' to the existing ONS LS for Wales, with the existing 1% Wales sample within the ONS LS being incorporated into a larger study. To help inform this decision, it would be useful to commission a project to examine the size of the Welsh sample within the existing ONS Longitudinal Study. As a 1% sample, it would be expected that the ONS LS would contain data on approximately 30,000 people currently living in Wales. Some of these

individuals will have been present in the LS since the extraction of the first sample based on 1971 Census data. It would also be useful to examine with ONS and CeLSIUS the feasibility and mechanisms required to undertake data linking activities with the existing ONS LS. Subject to necessary approvals, ONS and CeLSIUS are both keen to enhance and widen the use of the LS. Such 'test' exercises would provide a clearer idea of the functionality of such a database and would therefore be useful to inform decisions as to whether or not a full Census based Wales LS should be pursued.

Whilst additional resources would be required, creating a Census-based WLS within ONS would benefit from the existing research and governance infrastructure that exists to support the ONS LS and, in some respects, the WLS could be constructed at marginal cost. The recent availability of 2011 Census data (it is envisaged that the WLS could contain 2001 and 2011 Census data) would also make this decision a timely one. Although the ONS Longitudinal Study Development Team is based in Titchfield, the location of the ONS Head Office in Newport could provide a Wales based location where, subject to necessary oversight, the WLS could be accessed on site via the VML or other form of safe setting.

However, a number of issues need to be considered:

- Due to issues of disclosure, a 100% sample would not be feasible. A census based Wales Longitudinal Study that provided comparable levels of statistical power to the existing ONS LS and NILS would imply a sampling fraction of approximately 17%. It would be difficult to justify a larger sampling fraction than this for Wales.
- The construction of a WLS would require the cooperation of the HSCIC for England so that the tried and trusted technical and security infrastructure that has been used to successfully construct the ONS LS could be applied to the development of a WLS.
- It would be envisaged that SAIL would remain the primary repository and point of access for researchers wishing to analyse Welsh sources of administrative data where it was not necessary to merge socio-economic on to these data sets for the purpose of analysis. Furthermore, SAIL would also remain the primary location for research projects that required administrative data to be appended to surveys where consent to link had been sort. In both cases, Census data would not be required and existing mechanisms can be used to facilitate such links. Only in cases where Census data were required to be merged on to administrative sources would researchers be directed to the WLS.
- An important aspect of the WLS is the ability to link additional sources of administrative data regarding the population of Wales in to the database. These sources could be integrated in to the WLS on a project by project basis subject to the necessary approvals.

Such functionality would appear to be feasible given the ability of SAIL/NWIS to allocate NHS id numbers to administrative sources of data.

- Whilst it is envisaged that SAIL/NWIS could have a close working relationship with an ONS based WLS, both as suppliers of data and users of the WLS, an ONS based WLS would represent a separate and additional investment in data linking in Wales. However, whilst in some respects it would be preferable to have these activities 'under one roof', this arrangement would seem to be similar to that which exists in Scotland for SHIP and the SLS. It would also seem consistent with the approach of only providing researchers with the least amount of data necessary to conduct their analysis.
- Both the increased sample size and the functionality to link external sources of data may increase the risk of a WLS to such a level that existing infrastructure and access arrangements that are in place for the ONS LS are no longer deemed sufficient. Detailed discussions and development activities would need to take place in order to reach agreement surrounding the parameters of the WLS and the resources that would be required to establish such a database. Whilst ONS are willing to engage in such discussions, it cannot be assumed that ONS would ultimately support the construction of a WLS.

Whether it be in addition to or as alternative course of action to the establishment of an ONS Census based study, over the longer term to be able to secure a Census-based Welsh Longitudinal Study the Welsh Government may wish to give consideration as part of the Beyond 2011 programme as to whether it can seek access to Census data for Wales as part of the future solution or even whether it would wish to take responsibility for the Census in Wales. Given the expertise that exists at SAIL in terms of linking administrative data, NISCHR's ongoing commitment to making best use of routine data for research and the devolution of powers to the Welsh Government in the area of Health, it is difficult to imagine that Wales would not already have its own Longitudinal Study if it were not for the absence of devolved responsibility for the Census in Wales. The current position in Wales appears to be inequitable. Subject to the recommendations of the Beyond 2011 Programme, the 2011 Census is likely to be the last of its type. Although there is uncertainty surrounding the future structure of the Census, this may be an opportune time to reflect upon these issues.

If an ONS based solution is not regarded as feasible, the opportunities for developing a WLS appear limited. SAIL incorporates data from the Welsh Demographic Service, an alternative source of data that could be used as a population spine through which other data sets can be linked. Whilst previous research has demonstrated that such data does over-estimate the size of the total population, findings from the Beyond 2011 Programme reveal it to be a good quality

source of population data. The problem with this data is that it only contains information on gender, age and geographical location; no socio-economic data is available. Whilst there is nothing to prevent survey and administrative sources being held together in a database by this population spine, such a database would not contain socio-economic data for all participants; one of the key features of the existing Longitudinal Studies. This precludes its use for understanding how outcomes vary between different population sub-groups and to support statistical matching exercises. The emphasis of such a database will be more upon discrete data linking activities, such as enhancing existing survey data or developing links between administrative sources. It is interesting to note that each of the options for the Beyond 2011 Programme therefore incorporate a survey or traditional census element to overcome the shortcomings of administrative data.

The main sources of individual level socio-economic data that are currently (or will shortly become) available through SAIL are those which can be derived from relatively small cross-sectional surveys conducted by the Welsh Government where respondents have given their informed consent for data linking. More attention should be given to the development of a set of harmonised core questions that would be required to appear in all Welsh Government surveys so that the feasibility of pooling data across different sources is enhanced. The ability of the Annual Population Survey, Wales' largest regular household survey, to contribute to the WLS also appears to be limited. As an ONS survey, similar difficulties emerge with respect to the APS to those that emerge with respect to the Census. It is difficult to see how consent to link questions could be incorporated on to the APS or how the arrangement between WG and ONS could be renegotiated to allow WG to have greater freedom over how APS data can be used. Nonetheless, the APS remains an important underutilised source of longitudinal data for Wales.

Whether or not it ultimately proves feasible to construct a Wales Longitudinal Study, this should not detract from the importance of continuing to identify new sources of administrative data that can be deposited in SAIL and continuing to include consent to link questions in WG surveys so that new opportunities for research become available. The benefits of identifying and depositing such sources of data will be enhanced further if the recommendations of the recent report of the Administrative Data Taskforce are implemented; specifically those related to the establishment of an Administrative Data Research Centre in each of the four countries of the UK and the establishment of a UK Administrative Data Research Network that will responsible for linking data between government departments. Such developments may pave the way for data that is held to be linked to a far wider range of administrative sources (such as employment data held by HMRC and DWP). Although not as rich as Census data, these sources may, to some degree, provide an alternative way of combining more socio-economic data with existing WG sources.

The growth in usage of research data centres, such as the VML, the Secure Data Service and the DSUs that support access to the 3 existing UK Longitudinal Studies tends to be organic. The

analysis of new administrative data sets involves, or can be perceived to involve, considerable start up costs which may be off-putting to researchers who are used to analysing more well established sources of secondary data. Once new sources of linkable survey data or administrative data are available in SAIL, the Welsh Government should consider the use of small research grants (such as its Economic Research Grant scheme) in order support the use of these data sets. Attention should also be given to developing a user-support functionality of SAIL so that access to data sources at SAIL is free at the point of use. This would enable academics who do not have dedicated funds to conduct secondary analysis on data deposited at SAIL.

Annex 1: Patterns of Appearances in the APS Panel Database

Years Present							Main LFS Sample				Enhanced Sample			
04	05	06	07	08	09	10	England	Wales	Scot	North Ireland	England	Wales	Scot	Total
0	0	0	0	0	0	1	89,901	5,246	9,801	4,280	43,885	11,778	12,631	177,522
0	0	0	0	0	1	0	46,730	2,601	5,037	2,096	18,123	4,689	5,089	84,365
0	0	0	0	0	1	1	45,731	2,550	4,859	2,409	23,510	6,897	7,760	93,716
0	0	0	0	1	0	0	42,623	2,437	4,461	1,967	16,032	3,683	4,356	75,559
0	0	0	0	1	0	1	0	0	0	0	3,497	820	1,096	5,413
0	0	0	0	1	1	0	54,424	3,222	5,718	2,779	7,841	1,856	2,143	77,983
0	0	0	0	1	1	1	0	0	0	0	15,406	4,153	5,449	25,008
0	0	0	1	0	0	0	44,143	2,482	4,499	2,076	14,723	3,933	4,486	76,342
0	0	0	1	0	0	1	0	0	0	0	440	121	138	699
0	0	0	1	0	1	0	0	0	0	0	1,406	367	490	2,263
0	0	0	1	0	1	1	0	0	0	0	1,261	433	453	2,147
0	0	0	1	1	0	0	55,942	3,513	5,673	2,740	6,520	1,899	2,237	78,524
0	0	0	1	1	0	1	0	0	0	0	1,576	433	562	2,571
0	0	0	1	1	1	0	0	0	0	0	4,509	1,386	1,520	7,415
0	0	0	1	1	1	1	0	0	0	0	10,387	3,414	3,977	17,778
0	0	1	0	0	0	0	42,975	2,472	4,212	2,188	16,764	3,895	4,339	76,845
0	0	1	0	0	1	0	0	0	0	0	413	84	99	596
0	0	1	0	1	0	0	0	0	0	0	1,332	344	410	2,086
0	0	1	0	1	1	0	0	0	0	0	1,042	296	376	1,714
0	0	1	1	0	0	0	56,387	3,564	5,852	2,739	6,604	1,741	2,128	79,015
0	0	1	1	0	1	0	0	0	0	0	1,285	405	455	2,145
0	0	1	1	1	0	0	0	0	0	0	4,955	1,519	1,752	8,226
0	0	1	1	1	1	0	0	0	0	0	11,051	3,401	3,848	18,300
0	1	0	0	0	0	0	45,898	2,624	4,563	2,076	34,089	5,007	5,447	99,704
0	1	0	0	1	0	0	0	0	0	0	343	84	83	510
0	1	0	1	0	0	0	0	0	0	0	1,065	315	388	1,768
0	1	0	1	1	0	0	0	0	0	0	1,067	302	399	1,768
0	1	1	0	0	0	0	57,894	3,584	6,135	2,764	17,429	2,360	2,932	93,098
0	1	1	0	1	0	0	0	0	0	0	1,268	385	502	2,155
0	1	1	1	0	0	0	0	0	0	0	5,001	1,577	2,066	8,644
0	1	1	1	1	0	0	0	0	0	0	11,476	4,023	5,211	20,710
1	0	0	0	0	0	0	104,604	6,539	11,033	5,212	67,848	14,635	12,056	221,927
1	0	0	1	0	0	0	0	0	0	0	287	62	108	457
1	0	1	0	0	0	0	0	0	0	0	1,278	713	929	2,920
1	0	1	1	0	0	0	0	0	0	0	1,091	313	470	1,874
1	1	0	0	0	0	0	56,620	3,389	6,454	2,855	11,241	9,372	8,455	98,386
1	1	0	1	0	0	0	0	0	0	0	1,107	297	413	1,817
1	1	1	0	0	0	0	0	0	0	0	7,138	5,854	6,794	19,786
1	1	1	1	0	0	0	0	0	0	0	11,091	3,762	4,898	19,751
							743,872	44,223	78,297	36,181	385,381	106,608	116,945	1,511,507

Annex 2: Patient Register Population Estimates

Table A2.1 ONS and Patient Register Population Estimates: Ages 1-4

	Males			Females		
Ages 1-4	**GP Register**	**MYPE**	**GP/ONS Differential**	**GP Register**	**MYPE**	**GP/ONS Differential**
Isle of Anglesey	1523	1539	-1.1%	1427	1423	0.2%
Gwynedd	2585	2478	4.3%	2562	2542	0.8%
Conwy	2315	2300	0.7%	2233	2190	1.9%
Denbighshire	2253	2194	2.7%	1993	1956	1.9%
Flintshire	3620	3512	3.1%	3439	3429	0.3%
Wrexham	3364	3332	0.9%	3209	3187	0.7%
Powys	2670	2650	0.7%	2564	2531	1.3%
Ceredigion	1328	1372	-3.2%	1259	1241	1.5%
Pembrokeshire	2669	2679	-0.4%	2485	2491	-0.2%
Carmarthenshire	4039	3998	1.0%	3870	3837	0.8%
Swansea	5522	5333	3.5%	5046	4862	3.8%
Neath Port Talbot	3221	3218	0.1%	3115	3089	0.8%
Bridgend	3305	3281	0.7%	2996	3034	-1.2%
Vale of Glamorgan	2907	2940	-1.1%	2854	2813	1.4%
Rhondda, Cynon	5731	5699	0.6%	5554	5478	1.4%
Merthyr Tydfil	1489	1502	-0.9%	1344	1295	3.7%
Caerphilly	4337	4339	0.0%	4159	4088	1.7%
Blaenau Gwent	1533	1513	1.3%	1586	1554	2.1%
Torfaen	2226	2208	0.8%	2135	2138	-0.1%
Monmouthshire	1920	1839	4.4%	1767	1767	0.0%
Newport	3700	3569	3.7%	3637	3541	2.7%
Cardiff	8814	8211	7.3%	8416	7962	5.7%

Table A2.2 ONS and Patient Register Population Estimates: Ages 5-7

	Males			Females		
Ages 5-7	GP Register	MYPE	GP/ONS Differential	GP Register	MYPE	GP/ONS Differential
Isle of Anglesey	1107	1133	-2.3%	1030	1022	0.8%
Gwynedd	1915	1820	5.2%	1689	1622	4.1%
Conwy	1724	1708	0.9%	1590	1590	0.0%
Denbighshire	1583	1531	3.4%	1411	1397	1.0%
Flintshire	2590	2570	0.8%	2465	2443	0.9%
Wrexham	2274	2157	5.4%	2168	2132	1.7%
Powys	2055	2041	0.7%	1924	1911	0.7%
Ceredigion	1019	974	4.6%	960	901	6.5%
Pembrokeshire	1866	1848	1.0%	1820	1805	0.8%
Carmarthenshire	3015	2991	0.8%	2733	2734	-0.1%
Swansea	3927	3758	4.5%	3655	3486	4.8%
Neath Port Talbot	2147	2149	-0.1%	2170	2141	1.4%
Bridgend	2321	2284	1.6%	2268	2261	0.3%
Vale of Glamorgan	2099	2087	0.6%	2003	1992	0.5%
Rhondda, Cynon	4071	3983	2.2%	3903	3844	1.5%
Merthyr Tydfil	1016	967	5.0%	922	902	2.2%
Caerphilly	3267	3248	0.6%	2888	2874	0.5%
Blaenau Gwent	1110	1041	6.6%	1083	1024	5.8%
Torfaen	1511	1515	-0.3%	1435	1407	2.0%
Monmouthshire	1483	1498	-1.0%	1386	1344	3.1%
Newport	2663	2525	5.5%	2540	2427	4.7%
Cardiff	5862	5246	11.7%	5624	4964	13.3%

Table A2.3 ONS and Patient Register Population Estimates: Ages 6-11

	Males			Females		
Ages 8-11	**GP Register**	**MYPE**	**GP/ONS Differential**	**GP Register**	**MYPE**	**GP/ONS Differential**
Isle of Anglesey	1539	1568	-1.9%	1464	1524	-3.9%
Gwynedd	2757	2834	-2.7%	2610	2629	-0.7%
Conwy	2520	2397	5.1%	2524	2384	5.9%
Denbighshire	2311	2332	-0.9%	2225	2152	3.4%
Flintshire	3664	3617	1.3%	3484	3434	1.5%
Wrexham	3261	3193	2.1%	3048	2983	2.2%
Powys	3071	3061	0.3%	2873	2756	4.2%
Ceredigion	1552	1573	-1.3%	1422	1458	-2.5%
Pembrokeshire	2936	3039	-3.4%	2633	2694	-2.3%
Carmarthenshire	4176	4180	-0.1%	3918	3840	2.0%
Swansea	5432	4855	11.9%	5178	4856	6.6%
Neath Port Talbot	3295	3146	4.7%	3007	2900	3.7%
Bridgend	3321	3163	5.0%	3263	3107	5.0%
Vale of Glamorgan	3108	3184	-2.4%	3004	3039	-1.2%
Rhondda, Cynon	5636	5548	1.6%	5366	5309	1.1%
Merthyr Tydfil	1402	1303	7.6%	1312	1205	8.9%
Caerphilly	4405	4327	1.8%	4092	3997	2.4%
Blaenau Gwent	1580	1572	0.5%	1565	1531	2.2%
Torfaen	2226	2106	5.7%	2100	2040	2.9%
Monmouthshire	2063	1939	6.4%	2073	1951	6.3%
Newport	3813	3637	4.8%	3600	3514	2.4%
Cardiff	7994	7132	12.1%	7742	6827	13.4%

Table A2.4 ONS and Patient Register Population Estimates: Ages 12-15

	Males			Females		
Ages 12-15	**GP Register**	**MYPE**	**GP/ONS Differential**	**GP Register**	**MYPE**	**GP/ONS Differential**
Isle of Anglesey	1732	1794	-3.5%	1523	1544	-1.4%
Gwynedd	2798	2822	-0.9%	2831	2883	-1.8%
Conwy	2977	2650	12.3%	2835	2566	10.5%
Denbighshire	2466	2352	4.8%	2442	2312	5.6%
Flintshire	3917	3814	2.7%	3870	3754	3.1%
Wrexham	3403	3262	4.3%	3073	3082	-0.3%
Powys	3460	3499	-1.1%	3240	3204	1.1%
Ceredigion	1732	1800	-3.8%	1554	1606	-3.3%
Pembrokeshire	3034	3083	-1.6%	2925	2998	-2.5%
Carmarthenshire	4557	4655	-2.1%	4396	4411	-0.4%
Swansea	5694	5205	9.4%	5378	4867	10.5%
Neath Port Talbot	3610	3459	4.4%	3433	3295	4.2%
Bridgend	3675	3431	7.1%	3386	3218	5.2%
Vale of Glamorgan	3354	3384	-0.9%	3251	3294	-1.3%
Rhondda, Cynon	6110	5966	2.4%	5731	5587	2.6%
Merthyr Tydfil	1555	1407	10.5%	1584	1474	7.5%
Caerphilly	4790	4550	5.3%	4559	4423	3.1%
Blaenau Gwent	1797	1807	-0.6%	1749	1772	-1.3%
Torfaen	2560	2443	4.8%	2418	2334	3.6%
Monmouthshire	2587	2465	4.9%	2324	2254	3.1%
Newport	4065	3867	5.1%	3774	3639	3.7%
Cardiff	8558	7863	8.8%	8047	7204	11.7%

Annex 3: Characteristics of LFS/HBAI Samples (2008/10)

	LFS	HBAI
Age band		
16 to 24 yrs	3.1	3.3
25 to 34 yrs	12.6	13.4
35 to 44 yrs	17.9	17.3
45 to 54 yrs	18.4	14.0
55 to 64 yrs	18.8	18.9
65 to 74 yrs	16.7	17.4
75 to 79 yrs	5.4	6.2
80+ yrs	7.0	9.5
Family type		
Male pensioner	4.4	6.1
Female pensioner	12.1	13.6
Pensioner couple	19.0	21.5
Couple with 1 child	8.6	7.6
Couple with 2 children	9.1	8.9
Couple with 3 children	3.9	4.2
Couple no children	18.2	14.5
Lone parent with 1 child	4.5	3.8
Lone parent with 2 children	3.2	3.0
Lone parent with 3 children	1.2	1.3
Single male	8.5	9.6
Single female	7.4	6.0
Economic Status		
Self employed	11.9	8.4
All working full time	31.5	22.7
One full, one part time	13.7	7.0
One full, one non-working	5.9	6.7
One or both part time	11.5	7.4
Workless pensioners	20.8	35.8
Workless, unemployed	0.6	2.4
Workless, inactive	4.2	9.7
Housing Tenure		
Owns outright	37.7	39.3
Own with mortgage	33.9	30.1
Part own part rent	0.2	0.1
Rent council	10.3	11.5
Rent local authority	5.7	7.4
Rent other	11.2	10.4
Rent free	1.0	1.3
Total	100	100
Sample	6,644	2,968