School for Social Care Research

NHS
National Institute for Health Research

Methods Review 5

The use of 'large scale datasets' in UK social care research

Shereen Hussein

Improving the evidence base for adult social care practice

The School for Social Care Research

The School for Social Care Research is a partnership between the London School of Economics and Political Science, King's College London and the Universities of Kent, Manchester and York, and is part of the National Institute for Health Research (NIHR) http://www.nihr.ac.uk/.

The School was set up by the NIHR to develop and improve the evidence base for adult social care practice in England. It conducts and commissions high-quality research.

About the author

Shereen Hussein, BSc MSc PhD, is a Senior Research Fellow at the Social Care Workforce Research Unit, King's College London. Over the past decade, Shereen has been working in the domain of long-term care and policy-related context in the UK and for the World Bank. Her current research includes migration and long-term care, the effect of workforce structure and dynamics on workers' stress and intention to leave, adult abuse and safeguarding and social work models. With a strong background in statistics and demography, she is particularly interested in statistical modeling of large datasets to inform different policy questions and debates and she is the author of the Social Care Workforce Periodical. Shereen has previously worked internationally for the United Nations and the Population Council conducting research in the Middle East and North Africa on child morbidity, women's status and family formation.

NIHR School for Social Care Research
London School of Economics and Political Science
Houghton Street
London
WC2A 2AE

Email: sscr@lse.ac.uk
Tel: +44 (0)20 7955 6238
Website: www.sscr.nihr.ac.uk

ISBN 978-0-85328-450-5

ABSTRACT

This methods review sets out knowledge about current uses and applications of large datasets for research in adult social care practice. Built on a wide-ranging search of the literature, this review discusses examples of the use of different large datasets such as the General Social Care Council, the Census, the Labour Force Survey, governmental and hospital records, as well as others in health and social care research. It focuses on the methods adopted to extract and use data from different large datasets to enable quantitative and statistical examination of the research questions considered. It discusses the challenges associated with using large data records, which in some cases are not originally designed for specific quantitative data analysis or to answer a pre-defined research question, and illustrates various approaches adopted by researchers to extract, validate, refine and interpret results based on such data. The review discusses the strengths and limitations of a number of large datasets currently used in research on social care in England, with examples from adult safeguarding and other areas relevant to adult social care practice.

RECOMMENDATIONS FOR RESEARCH ON ADULT SOCIAL CARE PRACTICE

This review recommends that social care researchers:

- Accredit the potential value of existing datasets that are relevant to social care practices.
- Invest in identifying available datasets that are relevant to their research questions to produce more robust research findings. This may include combining or benefiting from different datasets.
- Recognise the value of adopting appropriate statistical methodologies when analysing existing large datasets within a multi-methods research design.
- Address the challenges associated with utilising existing datasets when examining new research questions.

KEYWORDS

Large datasets, quantitative analysis, administrative data, social care, statistical modelling, research methods

ACKNOWLEDGEMENTS

I am very grateful to the NIHR School for Social Care Research for commissioning this review, with particular thanks to Dr Michael Clark and Anji Mehta for their support. I am especially grateful to the anonymous reviewers, and to Professor Martin Knapp and Professor Jill Manthorpe for their valuable and constructive comments on an earlier draft of this overview.

CONTENTS

Introduction and background 1

Methods 2

Findings 2

- Types of available large datasets
- Strengths and considerations when analysing large datasets in the social care field
- Accessing data

Analysis 16

Discussion and conclusion 19

Recommendations 21

References 22

INTRODUCTION AND BACKGROUND

Secondary data analysis, defined as the use of existing data to address new research questions or methods (Pollack 1999), forms a cornerstone of research. Large datasets usually refer to existing data in all, or most, cases with specific characteristics; or to a large enough sample representing either a specific group or the general population. Secondary data usually include records at the individual 'micro-level', such as patient records, but may also refer to 'macro-level' records such as those related to services. The use of existing large datasets to examine, estimate, investigate and predict specific research questions is common practice in a number of disciplines including health-related research. National surveys and the census provide large samples of data offering information on a variety of topics including education, health and demographic factors, among many others. Increased technological capabilities in terms of storing and accessing large databases for different purposes have the potential to result in the availability of huge datasets, applicable to different disciplines (Berger and Berger 2004).

However, across UK social care research, the use of existing large datasets appears to be still in its infancy; although awareness of the potential of these sources to investigate new research questions is growing. This may be in part due to the shortage of large datasets that include elements specific to social care. Thus, it is important to highlight the need for large-scale surveys, or survey modules and questions, which are specific to social care. As Sharland (2009) in her strategic advice to the social work and social care research communities highlights:

> Valuable opportunities are being lost… both through lack of social work and social care input [into these datasets] and lack of use of these datasets to explore changing patterns and outcomes of social care problems and interventions, and relatedness of these to wider social and developmental change (pp. 13–14).

Large datasets have many clear advantages but also pose a number of challenges to researchers, both when adapting research questions and when choosing appropriate statistical methods. Datasets range from national coverage of all populations to samples of those with specific characteristics, and include users' records, patients' records, employee data or student databases. Routinely collected data, such as those collected by providers and commissioners, offer another useful set of information. The advantages of existing databases include the fact that they are readily available, and often free of charge; they are a potentially rich source of information about large numbers of people; and using existing data is generally less demanding (and has fewer ethical constraints) than planning, funding and executing long-term experimental studies. However, it is equally important to realise that most existing databases are often compiled for other purposes than social care research and may be observational rather than experimental.

The aim of this review is to provide a balanced view of these two perspectives, particularly within the context of social care research. The review uses examples of recent research

which used large datasets either solely or as part of a mixed method approach to examine research questions related to social care in the UK.

METHODS

This review is based on a literature search of journal and web-published materials using a search term strategy related to social care and health. The following databases were included: Applied Social Science Index and Abstracts (ASSIA); Health Management Information Consortium Database (HMIC); Social Care Online (SCO); Sociological Abstracts and Social Services Abstracts (SSA/SA). Other examples are drawn from nursing research through Medline, Pubmed and CINAHL (a nursing database) as well as Google Scholar. This is because much nursing research covers areas of social care such as care homes, as well as touching upon related areas of practice. The review also draws on personal experience of using a number of different large datasets specifically related to social care and social work. The datasets included in this review cover those held by Government departments and regulatory bodies, census and national survey data (such as the Labour Force Survey) and the emerging National Minimum Data Set for Social Care (NMDS-SC).

An initial search, restricted to the last ten years, retrieved over 500 articles from across the globe reporting on research that has analysed large datasets in the domains of nursing, social care and social work. Secondary data analysis of existing large datasets was either the main focus of the research or constituted an important element of the research. However, restricting the search to the UK nursing or social care (including social work) fields resulted in 280 articles. Further restricting the search to social care or social work in the UK led to less than 30 references being identified, with many focusing on discussing methodological issues of analysing large datasets. These mainly took place in the last five years, confirming the findings reported by Sharland (2009).

FINDINGS

In this section, some examples of using different existing large dataset sources and some of their applications in the social care research field are provided. Advantages and considerations when using existing datasets are critically discussed, with attention given to both methodological and analytical considerations.

Types of available large dataset

Census and national surveys data

The use of Census and national surveys data, such as the Labour Force Survey (LFS) and the British Household Panel Survey (BHPS), seems to be more common than the use of other specific datasets such as patient records, Primary Care Trusts' user satisfaction surveys and government databases. A number of general population-based databases such as the

Census and Labour Force Survey are used extensively by social scientists but to a lesser degree by researchers focusing on social care and social work. This seems to be an area for development because these large datasets provide researchers with ready-to-use data that usually cover good-size samples.

A census is a count of all people and households in the country. It provides population statistics from a national to neighbourhood level for government, local authorities, business and communities. Censuses are conducted every ten years and the last census for England and Wales has recently been completed (reporting on the night of 27 March and closing on 22 August 2011). This is the 21st full national census of the population. It is anticipated that it will have involved around 25 million households. The Census provides information on a number of important aspects of the whole population. Distinctive sections collect detailed data on population characteristics, health, housing, employment and transport. However, there are elements of the Census that are limited when answering some research questions, such as those relating to sexuality (Price 2011).

In addition to researching relationships and associations between different factors among the whole population, a sub-sample with specific characteristics can be identified for further research. For example, focusing on 'informal' care provision, Del Bono and colleagues (2009) used an individual Sample of Anonymised Records (SAR) of the UK 2001 Census data to identify the role of gender and partnership status in caring commitments for older people. Employing a set of logistic regression models (see below), they concluded that, after adjusting for marital status and household size, older women were significantly more likely to provide care than older men.

Using longitudinal records identified through Census data, McCann and colleagues (2009) focused on a sample of people aged 65 years and over at the time of the 2001 Census who were residents of care homes as a base for their longitudinal cohort study of variations in survival amongst residents of nursing and residential (care) homes in Northern Ireland. They employed a method of survival analysis techniques, namely a Cox proportional hazards model, showing that, after adjustment for age, sex, self-reported health (as a proxy of health status) and marital status, mortality (death) risk was 1.7 times greater in residential homes, 2.6 times higher in dual registered homes and 2.9 times higher for nursing home residents than equivalently aged people living in the community.

Linked to the Census is the Office for National Statistics Longitudinal Study (LS) from 1971–2001. The LS links Census and vital events information for one per cent of the population from 1971 to 1981, although for the latest 2001 Census only a provisional linkage is available. Young and Grundy (2008) examined possible associations between employment history, marital status and informal care provision among people aged 40–59 using these longitudinal linked data as constructed through the progressive linkage of vital events and the 1971 Census. Their findings, based on a logistic regression model, suggested a continuing gender dimension in informal (unpaid) care provision that interacts with marital status and employment.

Examination of other aspects of informal care has dominated the use of Census data in the social care research field; for example, Dahlberg *et al.* (2007) examined the effect of gender and age using logistic regression models and Doran *et al*. (2003) explored the health of young and older informal carers through the use of descriptive statistics.

The Labour Force Survey (LFS) is a quarterly sample survey of households living at private addresses in Great Britain. Its purpose is to provide information on the UK labour market that can then be used to develop, manage, evaluate and report on labour market policies. Use of the LFS within the social care field is usually an attempt to estimate the size of the social care workforce (for example, Eborall and Griffiths 2008) or sections within it, for example, the contribution of migrants to the health and social care sectors (Dobson and Salt 2009).

The British Household Panel Survey (BHPS) began in 1991 and is a multi-purpose study. It follows the same representative sample of individuals over a period of years. BHPS is household-based, interviewing every adult member of sampled households. Focusing also on workforce perspectives and concerning the supply of 'informal' caregivers, Carmichael and colleagues (2010) used 15 waves of the BHPS to examine the causal effect of employment status and willingness to provide informal care. Using longitudinal panel data allowed the researchers to identify in a given year, people who, while not yet carers, were to become so in future years (as the datasets cover a total of 15 years from 1991 to 2005). They assumed that employment status prior to caring was exogenous (i.e. is affected by external factors) and investigated the relationship between employment status and transitions into caring through the use of a discrete-time model. Broadly related to the social care field, using data from both the BHPS and the English Longitudinal Study of Ageing, Netuveli and colleagues (2006) examined the wider issue of quality of life in later life. Using a quality of life measure (CASP-19) and applying regression models, they found that quality of life was reduced by depression, poor perceived financial situation, limitations in mobility, difficulties with everyday activities, and limiting longstanding illness, while quality of life was improved by trusting relationships with family and friends, frequent contacts with friends, living in good neighbourhoods, and proxies of 'wealth' such as owing two cars.

At the other end of the age spectrum, a national survey of the mental health of children and adolescents in England and Wales was carried out on behalf of the Department of Health, the Scottish Health Executive and the National Assembly for Wales. The primary purpose of the survey was to produce prevalence rates of the three main childhood mental disorders: conduct disorder, hyperactivity and emotional disorders (and their co-morbidity). The second aim of the survey was to determine the impact and burden of children's mental health problems in terms of social impairment and adverse consequences for others. An additional purpose of the survey was to examine the use of services (both health and social care services). Examining service utilisation by children with conduct disorder (CD), Shivram and colleagues (2009) analysed data collected through the Great Britain Child Mental Health Survey – 2004 cohort. Their findings included a different

pattern of social services use by children with CD from those with emotional disorder. These large scale datasets may be of use to adult social care commissioners in indicating possible needs into adulthood.

Examining service utilisation among adults with mental health problems and whether 'age discrimination' exists, Beecham and colleagues (2008) analysed three nationally representative datasets as part of a multi-method research study. Using existing data from the Psychiatric Morbidity Survey 2000, longitudinal data from a randomised trial of treatments for people with depression and anxiety, and longitudinal data from an observational study of people with schizophrenia, they found an apparently reduced use of mental health services by older men compared to younger men but no differences among women. The analyses showed no age-cost differences between younger and older adults. However, using the two latter more specific datasets revealed some significant associations between age and both service use and cost.

The Health Survey for England is a series of annual surveys seeking information about the health of people living in England. It was commissioned by the Department of Health to provide better and more reliable information about various aspects of people's health and to monitor selected health targets. The survey combines questionnaire-based interviews with physical measurements and the analysis of blood samples. Blood pressure, height and weight, smoking, drinking and general health are covered every year.

Andrew (2005) analysed the Health Survey for England 2000 to investigate whether individual-level social capital is associated with care home residence and levels of mental and physical function among older adults. Using multivariate statistical analysis, he concluded that individual-level social capital is associated with both care home residence and a number of physical and mental health indicators. From the neighbouring health field, Gill and colleagues (2009) used data from the same survey for the years 1999 and 2004 to estimate the number of people requiring English language support from Black and Minority Ethnic (BME) groups in England. They combined data from two national surveys taking place in 1999 and 2004; both contained boosted samples of ethnic minority groups. Based on statistics obtained from these surveys, they estimated the size of different ethnic groups who are unable to speak English and may require interpreting services. They found the needs for such services increased with age and were higher among women from BME communities than men. These findings are clearly of relevance to social care providers.

Kavanagh and Knapp (2002) used the Office of Population Censuses and Surveys (OPCS) disability surveys conducted in the mid-1980s. The OPCS involved separate national surveys of disability among adults in communal establishments and private households, Kavanagh and Knapp examined the links between costs and levels of cognitive disabilities, using multivariate regression methods. The UK700 case management, randomised control trial was funded by the English Department of Health and NHS Research and Development Programme. It aimed to investigate the cost-effectiveness of intensive compared with

standard case management for patients with severe psychosis. It involved randomly allocating 708 patients with psychosis and a history of repeated hospital admissions to standard case management (case-loads 30–35) or intensive case management (case-loads 10–15). Clinical and resource use data were assessed over two years. This trial produced rich information and detailed data about the patients and the way cases were managed. A number of researchers reanalysed the data to examine different research questions. The UK700 group (2000) initially examined the cost-effectiveness of intensive versus standard case management for severe psychotic illness, using standard t-test mean comparison methods, with an additional boot-strapping exercise to validate the results. Huxley and colleagues (2001) used the same dataset to examine differences in quality of life (using QOL outcome measure) according to case management and diagnoses. They concluded that QOL was not associated with either factor; however, they suggested a better subjective QOL measure which was more sensitive in assessing improvements in depression. In 2001, Hassiotis and colleagues used the same data to investigate the outcome and costs of care among psychotic patients with borderline IQ relative to those with normal IQ levels. Intensive case management was significantly more beneficial for borderline-IQ patients. In the case of the UK700 trial data, there is a high level of consistency between the original purpose and design of the data and research questions being investigated, which is usually the case with very specifically designed and purposively collected datasets. These types of datasets are of great advantage when particular information and research questions related to a certain group are required by social care research.

Administrative and government department databases

In addition to the main Census and national survey data, a number of specific administrative and governmental databases are available and may be a rich resource for social care researchers. Some examples of such databases and their use in research are provided here.

Over the past few years the collection of data from adult social care users and carers – with an emphasis on the importance of users' and carers' satisfaction and perceived impact of social care services on their quality of life – has developed. Two major surveys are being introduced in England and are expected to provide researchers, for the first time, with large-sample data on social care users' and carers' experiences. The first survey is the *Adult Social Care Survey* (ASCS), an annual national survey with first collection taking place in 2010–11. The aim of the survey, which is collected from social care users by local authorities in England, is to improve the understanding of users' experiences and perceptions of the services received. The questionnaires are available in several languages to facilitate wider participation and incorporate a set of questions from the Adult Social Care Outcomes Toolkit (ASCOT). Each local authority, using a pre-agreed formula, selects a stratified sample of users to complete the survey. It is envisaged that the ASCS will provide much-needed information on personal outcomes for adults receiving social care (The NHS Information Centre 2011). The second national survey is the biannual *Carers' Experience Survey* (CES) which is due to start collecting data in 2012–13 from local authorities in

England. The survey is developed by the Personal Social Services Research Unit (PSSRU), and builds on previous PSSRU users' surveys conducted by 90 local authorities on a voluntary basis in 2009–10. The CES is a self-completed survey collecting information on carers' experiences and levels of satisfaction with services, incorporating a quality of life outcome measure (Fox *et al*. 2010).

The *Count Me In Census* is a national census that collects information on all inpatients in mental health services, including learning disability health services, with the clear aim of ensuring a high level of ethnic assessment and record keeping. The 2005 *Count Me In Census* was complemented by a service user survey aiming to obtain information on service users' experiences. The 'census' has been conducted yearly from 2005 to 2010. The literature search did not indicate any secondary data analyses of this database, possibly due to its recent appearance. However, some have argued in favour of its potential, particularly in identifying mental health needs among older people (for example, Shah 2009). Again, bearing in mind the extent of social work involvement with many service users surveyed by this census, its findings have relevance for social care providers and commissioners.

The *Children In Need Census* (CIN) is a yearly census conducted by the Department for Education with all local authorities in England and provides data on all cases and episodes involving children in need. Its 'Children Looked After' statistics are based on a return supplied by local authorities called the SSDA903, which uses a one-third sample of all children who have been looked after at any point during the year. Dickens and colleagues (2007) used these two data sources as part of their examination of inter-authority variation and case-centred decision-making. Vostanis and colleagues (2008) investigated service use by looked after children with behavioural problems by analysing existing data collected as part of the survey of looked after children. This survey was conducted with a purposively selected sample from English local authorities' databases of looked after children. The researchers briefly discussed the limitation of using this specific database, particularly in relation to the sample size after including an eligible group for their study. Other sources of child protection data have also been used; for example, Pugh and Jones (2004) applied survival analysis techniques to identify patterns of variations in child protection practice using child protection registration records in Wales. This has implications for adult social care research in that all of these surveys relate to young people who, while they may be 'moving' into adult services in an administrative sense, are likely to have similar and long-lasting needs to those expressed during their time in care. Research is beginning to identify the importance of looking at child abuse when supporting people with problems in adulthood and so estimates of affected populations may help in planning and commissioning.

Combining published estimates on the prevalence of autism among different age groups and estimating accommodation arrangements for children with autism spectrum disorder and intellectual disabilities using children in need data, Knapp and colleagues (2009) estimated the economic costs of autism in the UK. They drew on service use data from

their own recent studies, and included opportunity costs of lost productivity in addition to other service costs. However, due to lack of data on informal care they were not able to estimate its cost. This study provides a fine example of utilising a number of datasets produced by different studies to investigate a new research question. Some of the datasets used were related to the evaluation of person-centered planning (Roberston *et al*. 2005); a matched-group study comparing costs and quality of life outcomes for adult with intellectual disabilities (Felce *et al*. 2008); a study focusing on the outcomes of a health screening programme (Cooper *et al*. 2007); a study investigating the prevalence of physical and mental illness among a group of learning disabled adults and data from a trial of the use of neuroleptics in adults with intellectual disabilities and challenging behaviour (Tyrer *et al*. 2008). On the subject of cost of social care for different service user groups, Curtis and Netten (2005) drew on a number of different data sources, such as the Employer Organisation's annual national Social Services Workforce Survey and service mapping data.

Other specific data records which have been used to examine new research questions in the adult social care field were identified through the search undertaken for this review. Sondhi and Huggins (2005), for example, examined the effectiveness of a social care model for arrest-referral schemes by analysing case records with probabilistic linkage to the National Drug Treatment Monitoring System database. These were combined with qualitative elements, including observations and semi-structured interviews. In another record-linking study, Morgan and colleague (2000) linked all in-patient admission data, out-patient appointments, attendances at accident and emergency departments and mortality for all the resident population in a health district. They additionally linked all such information to the learning disability register compiled by the local authority to identify people with learning disabilities. They analysed these data to describe the epidemiology of learning disability and examine the pattern of care provision. Their findings indicated a positive correlation between prevalence of learning disability and social deprivation, and significant variations in service use by people with learning disabilities when compared to other groups of the population. The implications of this for social care providers are numerous; for example, ensuring that carers' groups are easy to access in deprived areas.

In relation to adult protection or safeguarding, employers of social care staff working with vulnerable adults in England and Wales have been legally required to refer workers or volunteers dismissed for misconduct to the Protection of Vulnerable Adults (POVA) list because they have harmed vulnerable adults or placed them at risk of harm. From July 2004, it became a statutory requirement for care providers to check if new care workers/volunteers are included on the POVA list. This list was initially kept by the then Department for Education and Skills (DfES). More recently, from October 2009, the POVA scheme has been replaced by a new Vetting and Barring Scheme under the Safeguarding Vulnerable Groups Act 2006. Referrals are now made to the Independent Safeguarding Authority (ISA). Hussein and colleagues (2009a, 2009b, 2009c) analysed full records of referrals made to the former POVA list to examine a number of research questions in

combination with qualitative data sources. In addition they extracted further information from a sample of detailed records to enable them to examine additional relationships such as users' (alleged victims') characteristics and mitigations claimed by referred staff. This project examined associations between different staff characteristics and the probability of being barred after referral. The advantages and challenges of using this particular database are discussed in detail by Hussein and colleagues (2010a). Briefly, for social care providers, this analysis indicated that there were very different reasons for referrals among workers in residential settings compared to home care settings. This suggests that different abuse-prevention strategies might be relevant to the different settings.

The National Minimum Data Set for Social Care (NMDS-SC) was the first attempt to gather standardized workforce information for the social care sector in England. It is developed, run and supported by the sector skills body, Skills for Care (SFC), and aims to gather a 'minimum' set of information about services and staff across all service user groups and sectors within the social care sector in England. In 2009, the Social Care Workforce Research Unit, King's College London, launched the Social Care Workforce Periodical (SCWP). SCWP is a regular web-based publication, where in each issue one aspect of the workforce is investigated through in-depth statistical analysis of up-to-date NMDS-SC workforce data to provide evidence-based information. By August 2011, thirteen issues had been web-published examining a number of workforce characteristics including: turnover and vacancy rates, demographic profile of the workforce, contribution of older workers; profile of younger workers, levels and variations in pay within the sector and contributions of migrants to the sector (see, for example, Hussein 2010a and Hussein 2011). The NMDS-SC, containing records on more than a quarter of a million social care workers, has also been used, either through additional quantitative analysis or as extracts from analysis provided by SFC, to examine the contribution of migrant care workers in England (Cangiano *et al*. 2009; Hussein *et al.* 2010b). Unlike the government datasets mentioned above, the NMDS-SC has limitations including missing data and currently being more representative of the independent sector than local authorities. However, it was decided in April 2011 after a national consultation that the NMDS-SC will be the central source of adult care workforce data for the sector, including local authorities, which is likely to increase its coverage substantially. It is important for large-scale dataset research to report limitations, such as the risk of bias. This is discussed further later in this review.

Social workers are required to register with the General Social Care Council before working in England. Three similar registers exist in the UK: in Wales – held by the Care Council for Wales; in Scotland – held by the Scottish Social Services Council; and in Northern Ireland – held by the Care Council for Northern Ireland. Such data records contain a number of important characteristics about social workers including country of training. Hussein and colleagues (2010b) used this dataset to examine the profile of and trends in non-UK qualified social workers in England as part of their migrant care workers research.

Other featured databases in research are records routinely collected about nursing and social work students and linking personal characteristics to their progression. Such records

have been used to examine students' characteristics such as age and ethnicity and rates of academic progression. For example, Mulholand and colleagues (2008) used descriptive and chi-square statistics to examine the association between nursing students' personal characteristics and their higher education progression. Hussein and colleagues (2009d) used multi-level modelling techniques to examine the hierarchical effects of higher education characteristics on one level and social work students' personal characteristics on individual levels on students' progression rates. The latter study focused on the new social work degree students and followed a previous study on social work students who were registered for the former diploma in social work (Hussein *et al.* 2008). The three studies, relating to nursing and social work students, showed significant association between diversity characteristics and students' progression rates. Routinely collected students data records were also analysed by the Evaluation of the Social Work Degree team (2008) as part of a wider study focusing on the transition of social work education into a degree level. Such studies are useful in calculating supply and demand for professions such as social work.

Strengths and considerations when analysing large datasets in the social care field

It is evident from the examples provided above that existing large datasets which are related to the social care sector offer a great deal of potential value to the research community and potentially therefore to users, practitioners, commissioners, providers and policy makers. National databases, as well as specific administrative and governmental databases, all feature in the literature. This review provides a glimpse into the spread and depth of existing datasets and how they can be utilised to examine new research questions or validate long-debated ones. Although the numbers of articles utilising existing large datasets in the social care field are not substantial, they illustrate a variety of different sources that have been used to examine a wide range of relevant research questions. However, the literature search for this review highlighted the fact that the use of such sources is still very limited when compared to, for example, the nursing field or with other countries, particularly the US.

Existing large datasets provide great advantages to researchers. However, they also require careful consideration and pose a number of challenges to both methods of analysis and in defining and agreeing the research questions considered. In this section of the review both the strengths and challenges of secondary analysis of large datasets are outlined.

One of the apparent advantages of reanalysing existing datasets is their actual existence as recorded information which usually relates to a large group of subjects. As considered above, coverage can vary from whole population censuses to specific records on a subgroup who may posses certain characteristics or fall within a certain group. Such databases already exist. Therefore the considerable research costs associated with questionnaire design, data collection and processing are avoided. Additionally, they usually provide data on a substantial sample size that is typically very difficult to achieve in most studies due to time and cost considerations.

There are growing opportunities, particularly in attempting to link different databases, to obtain more substantial information on larger groups of 'subjects'. In the social care field, as with other fields, there are increasing volumes of routinely recorded data with generally improving data quality. Technological progress has enabled large datasets to be processed, stored and shared (for example the ESRC Administrative Data Liaison Service (ADLS), which provides support to academics in identifying and accessing major administrative datasets in the UK). Existing datasets may contain hundreds of variables and attributes of great interest to social care researchers. The opportunities to test care-related theories, generate new knowledge for practice and evaluate different outcomes are numerous. With such great potential, analyzing large datasets is gaining international recognition as a valuable method in research.

For instance, recruiting 'hard-to-reach' groups (e.g. older people from minority ethnic groups) can be challenging. Therefore obtaining a large enough sample to compose reliable findings may be simply unattainable within the resources of an individual study. This was the case for the National Elder Abuse Study, which included over 2000 participants in a sample weighted to represent the general UK population aged 66 years or over living in the community (O'Keefe *et al*. 2007). Here, while a large survey took place, targeted sampling was necessary to achieve a relatively representative proportion of 2 per cent of participants from black and minority ethnic groups (the proportion is 2.5 per cent among the general UK population aged 66 or above).

Using existing large scale national surveys may offer information about a large enough sample of such hard-to-reach groups. Indeed, a number of research projects have used UK national surveys, such as the General Household Survey, to locate certain groups. Some of these surveys, for example the Family Resources Survey, are used to generate or 'host' specific modules for a sub-sample of the population (Sin 2006). Consequently, the use of existing large datasets in the social care research field may accelerate the pace of research by allowing researchers to ask complicated questions and gain more generalisable findings. Such evidence-based findings, obtained from large data analyses, can then guide further specific research questions. This is not to imply that secondary data analysis of large datasets is a quick venture. In fact, to conduct a valid and rigorous secondary data analysis, researchers need to pay considerable attention to a number of important factors such as the purpose of data collection, sampling design and the choice of adequate statistical modeling technique.

By definition, secondary data analysis, has fewer ethical concerns than primary research that includes planning, accessing and obtaining data from different groups of participants. However, when analysing large datasets, researchers should be clear about data ownership and maintaining anonymity of records. It is good practice to acknowledge the efforts of those who have undertaken the initial work or those who have been of particular assistance; in some cases there is also the need to seek permissions to reproduce extracts or tables.

Choosing the most suitable dataset

Locating a suitable database to examine for a particular research question is the starting point of using this approach. In some cases, there is only one dataset that contains the required information, such as when analyzing adult abuse referral cases to the POVA list (Hussein *et al*. 2009a). In other situations where information on attributes and people that are relevant to the research questions can be extracted from different sources, researchers need to apply some care in their choices. Some elements to be considered are the coverage of different datasets, the time frame of data collection, what other variables and attributes are available in each dataset, which may be theoretically linked to the research question and the quality of the data itself in terms of organization and coding. These are some of the key elements when deciding on a dataset. Other elements may include which dataset can provide an adequate reference group to compare to and how accessible, expensive and comparable the dataset is.

Using an existing dataset may limit researchers in terms of both the coverage of questions asked and the depth of data in relation to the particular questions under consideration. There have been greater efforts to include more detailed information on variables such as ethnicity and national identity in the Census and national surveys, which have helped in the investigation of a number of research questions related to ethnicity. However, other important personal characteristics, including sexuality, are still under-reported or not collected, as identified by Purdam and colleagues (2008). Only a few of the major UK surveys ask about sexual orientation (see Price 2011). Key surveys for measuring socioeconomic circumstances including the Census, the Labour Force Survey and the General Household Survey do not include a question on sexual orientation. These surveys tend to ask about the respondent's household and marital status, but same-sex couples are often treated as housemates whereas opposite-sex respondents living in the same house are treated as cohabiting. As part of the Office for National Statistics (ONS) harmonisation of questions on marital status, cohabiting as a couple can include same-sex couples.

The recent ONS consultation around the inclusion of a question on the 2011 UK Census identified a number of challenges of defining sexual orientation resulting in problems of data quality and accuracy. The ONS concluded that it was not possible to include such a question in the Census (ONS 2006). In the absence of a government census addressing issues of sexual orientation in the UK, ID Research set up the UK Gay and Lesbian Census in 2001 (ID Research 2002). Based partly on the findings from this and other research related to gay and lesbian people, a statement by the Public Administration Select Committee of the House of Commons (2009) recommended the inclusion of sexual orientation in the UK Census:

> The inclusion of a sexual orientation question is important in making the census relevant and useful in relation to equalities legislation and it is possible for the data to be sufficiently reliable (p. 10).

The 'rehearsal' Census questionnaires, which were administered in three local authorities in England on 11 October 2009, did not include any explicit questions related to 'sexual orientation'. However, another value was added to marital status options and now includes 'same-sex civil partners' as an option for marital status.

Another noticeable area where lack of specific information remains quite problematic is in the exact definition of what constitutes social care. Due to the imperfect match between data collection purposes and coverage, and the research question under investigation, researchers are usually compelled to use 'proxies' of important factors related to their research questions. For example, for many years researchers have analysed the Labour Force Survey data in attempts to estimate the social care workforce. However, due to lack of accurate definitions of the sector, findings always referred to the wider sector of 'health and social care'. The establishment of the National Minimum Data Set for Social Care (NMDS-SC) is a great step towards overcoming such gaps in knowledge. Indeed recent analyses (for example, those provided in the Social Care Workforce Periodical: http://www.kcl.ac.uk/scwru/pubs/periodical/) provide better opportunities to understand the structure and dynamics of the social care workforce. Other gaps in current datasets in relation to social care include finding suitable measures for quality and outcome of social care (see, for example, Malley and Netten 2008).

Record linkage

Record linkage and combining different databases provide great potential to explore a wider set of variables and attributes relating to the same 'subjects'. The linkage process of national data sources is usually the preserve of data-holding organisations such as the Office for National Statistics. Researchers may need to work in partnership with the data holders to perform such a process in some other situations, for example, when dealing with organisational data that are stored in different files and formats. This linkage process may pose particular challenges if there is no reliable unique identifier that can be used across cases. In the absence of 'unique' identifications that can be matched in two or more databases, record linkage is still possible. In such cases a number of variables are used to link the same person to data obtained from different datasets – for example using an algorithm to convert surnames and postcodes into a code to be combined to a transformation of date of birth. There are several approaches to record linkage, with the most straightforward referred to as 'a rules-based approach', in which reasonable rules are developed and then refined as common exceptions are found. The advantage of this approach is that it is possible to achieve considerable accuracy without needing a lot of labelled data to train or test the rules on. The disadvantage is that to obtain very high accuracy, more and more exceptions and special cases need to be handled, and eventually the list of rules gets too complex to be built by hand.

In addition to records linked through the National Office for Statistics - for example, linking Census and vital events information for one per cent of the population as used by Young and Grundy (2008) – other linkages on a smaller scale can be achieved. For example, in

research conducted by Sondhi and Huggins (2005) case records where an initial assessment in police custody had been undertaken were matched with data collected by the National Drug Treatment Monitoring System (NDTMS). They used a probabilistic matching approach. Probabilistic matching uses likelihood ratio theory to assign comparison outcomes to the correct, or most likely, decision. This method leverages statistical theory and data analysis and thus can establish more accurate links between records with more complex typographical errors and error patterns than deterministic systems. Typically, probabilistic systems assign a percentage (such as 90 per cent) indicating the probability of a match. Other matching approaches include deterministic matching systems, which may have a relatively lower degree of accuracy compared to probability matching.

Constructing longitudinal analyses

The recent availability of data within the adult social care sector invites constructing longitudinal analyses for the workforce, users and providers. There are huge advantages of analysing longitudinal data particularly if they relate to a large enough sample. In longitudinal data, information about a set of characteristics and variables are available for the same individuals (or subjects) at several time points. This longitudinal approach allows the investigation of individuals' changes over time. This is different from repeated cross-sectional data, which collects the same information about different groups of individuals (or subjects) at several time points. Statistically there are a number of advantages of the longitudinal approach – for example, subjects serve as their own controls, between-subject variation is excluded from error and it can provide more efficient estimators than cross-sectional designs with the same number and pattern of observations. The longitudinal design can also separate ageing (or time) effects from 'cohort' effects; this is not achievable from a cross-sectional design.

There are a number of challenges associated with the complexity of identifying and extracting longitudinal data from existing large datasets. There are also a number of challenges in relation to longitudinal data analysis; an overview of the latter set of challenges is provided here. Researchers need to be aware that observations are not, by definition, independent; therefore they need to account for dependency. There are also pragmatic problems in relation to availability of statistical software and modelling techniques that are specific to longitudinal design as well as computational requirements. In general, there needs to be a balanced approach taking account of the structure of longitudinal events, missing data and attrition.

One of the main datasets, particular for the care sector, which provides a potential for constructing longitudinal data is the National Minimum Data Set for Social Care (NMDS-SC). The NMDS-SC is updated regularly by a considerable number of social care providers in England. In each update the employers provide aggregate data on their workforce as well as detailed information on 50 to 100 per cent of their entire workforce. Skills for Care can provide unique identifiers for 'most' employees (workers) and the accumulative data provided by employers since the introduction of NMDS-SC in 2007. These information and

data records offer the opportunity to construct longitudinal analyses on individual as well as provisional levels. The process of collating and analysing longitudinal records has a number of challenges. Such challenges operate at different stages from dealing with very large data to using appropriate statistical models suitable for longitudinal structures. At the time of writing the Social Care Workforce Research Unit at King's College London is undertaking a major project focusing on identifying and utilising longitudinal data extracted from the NMDS-SC.

Another source of data is the historical records of social work students and data on the majority of the same group of previous students who are now registered social workers held by the General Social Care Council in England, and by other care councils in the UK. The construction of longitudinal data on the individuals spanning from being students and including their entry and movement with and outside the workforce can potentially provide a greater understanding of the profession's mobility and allow the construction of more reliable workforce supply models.

Conceptualising and mapping research questions to existing datasets

Reliable research findings will depend on using both data and statistical methods that are appropriate to the research question. The need for a conceptual framework that links and plots possible associations that are theory- and practice-based cannot be stressed enough, particularly when analysing large datasets as a core element of the research. Analyzing existing large datasets, or secondary data analysis, is a key component of applying this method of research. There are a number of issues to consider prior to deciding which statistical method to use. The first steps relate to some conceptual and methodological considerations. It is paramount to start the analysis with a valid theoretical framework of analysis or a conceptual model, which draws the relationships between different variables in a logical manner. Such a conceptual model helps to organise and classify existing data into useful structures, with a specific focus on issues that are relevant to the research question under consideration. Therefore, it is important to start by selecting the most suitable and relevant dataset that matches the research question being considered. However, sometimes general population surveys may contain very relevant and useful information. Thus, a thorough primary search of existing databases is recommended. Once a conceptual framework is developed and a suitable dataset is identified, large amounts of data may be used to test propositions related to a theory or to explore relationships between concepts of a theory. Some large datasets may contain a large number of variables and it is important to focus on the variables that relate to the conceptual framework and avoid the temptation of selecting unrelated, but possibly interesting, variables.

Given that existing datasets have been collected by people other than the researchers themselves, it is important to consider ways of minimising possible errors. After selecting a suitable database, it is advisable to consider what was the purpose of collecting such data, how the data were collected, how questions were asked, who responded to such questions, and how data were coded and stored. These are very important questions to

establish the accuracy and reliability of data. Firstly, did people answer for themselves or for others: was there a 'proxy' used? For example, in the NMDS-SC, data related to workers are reported by employers and not by workers themselves; therefore some personal information such as prevalence of disability may not be accurate. Another issue is that some datasets, in particular data which are routinely collected, are usually observational rather than experimental and contain 'outcome' data, such as vital events, but lack process data, for example 'how' a particular set of circumstances arose; this may impose certain interpretation limitations.

Accessing data

There has been recently growing interest in ustilising large datasets in the social care domain. Some of the main datasets related to the care sector can be identified and accessed through the National Adult Social Care Intelligence Services (NASCIS). The NHS Information Centre for health and social care holds data in relation to social care including carer support and user surveys. The Skills for Care for the National Minimum Data Set for Social Care and the NHS Information Centre Workforce provide workforce data. Social care and mental health indicators, from the National Indicator Set, can be downloaded from data.gov.uk; these may be linked to other datasets to provide macro information at a council level. Wider data sources can be obtained from the Office for National Statistics (ONS). It is also important to consider under-utilised organisational data and other governmental data that are relevant to the research question as described in the previous section.

ANALYSIS

It is important for researchers to realise that secondary data analysis does not eliminate the need for data management. In fact, some data sources (for example government data records) may require a great deal of data processing prior to analysis. The purpose and methods of data collection are key to establishing the level of data management. For example, if certain information is recorded without an adequate pre-coding strategy or using codes that are too broad for the research questions, a considerable amount of data management may be required before the analysis can start. Such processes are usually multi-step and quite elaborate. It is well acknowledged, and illustrated in research identified in this review, that there is no standard approach in coding administrative and governmental data related to social care. Therefore, it is important to allocate enough time for data exploration, mining and management, as well as familiarisation with the data and the variables they contain.

Another factor to be considered is the original sampling strategy used in collecting existing datasets. Many national surveys employ complex sampling designs, with an advantage that findings from these data are usually more generalisable than those obtained from smaller scale surveys. However, it is important that researchers be aware of

possible Type I and Type II errors (accepting a false association or rejecting a true association) and employ adequate statistical measures to account and identify them. Type I errors may result from 'noise' when there is no true effect, while Type II errors occur when the research fails to recognise a true effect. Several approaches can be implemented to reduce Type I errors, such as family-wise error (FEW) correction procedures and false discovery rate (FDR) techniques (Genovese *et al.* 2002). However, a main problem associated with attempts to only focus on reducing Type I errors is the direct cost in relation to increased Type II errors. The best estimate of Type II error rates comes from power analyses; these provide estimates of the likelihood of a Type II error in larger samples given the effect of a certain sample size.

Missing data are a concern in all types of research. It is important both to 'know' the large data analysed and to understand how missing data have been originally recorded. For example, the NMDS-SC record missing values as 'not available' and in some questions such as qualifications, both 'no qualifications' and 'missing' are coded as 'not available' which could be misleading in thinking about payment levels or training investment.

A number of considerations are therefore needed in selecting appropriate statistical analyses techniques and adequate software. For example, when analysing large survey data it is recommended to use special statistical modules (such as SUDAAN or svyset in STATA) that are particularly designed to adjust for sampling design and weights and will generally produce more accurate results. It is equally important to consider the hierarchical structure of the data, stratification and any weights used and to employ suitable statistical techniques and appropriate software. In some situations there is a need to consider the advice of software programmers when dealing with very large datasets. In reporting any results it is important to include a balanced description of data limitations and methods employed. A relatively small number of published articles reviewed here include sufficient descriptive information about the dataset used and its limitations (for example, Kavanagh and Knapp 2002).

After deciding on the data source, understanding its width and breadth, mapping existing variables to a conceptual framework and deciding which statistical software to use, an important element remains: the choice of the best statistical model or technique, which is suitable both for the data and for the research question. The literature search for this review reflected the dominance of multivariate analysis, particularly logistic regression models, in analysing social care research questions using large datasets (see for example Del Bono *et al*. 2009). There are a number of explanations for this phenomenon, mainly relating to non-normality assumptions of logistic regression models, the fact that most outcome variables of interest were binary and the relative ease of interpreting odds ratios that logistic regression models produce; for example, whether a person provides informal care or not (Del Bono *et al*. 2009), or whether a care worker has been previously working abroad or not (Hussein *et al*. 2010b). Given the considerable evidence of use of such a model it is worth briefly explaining its aim and purpose.

Binary (or binomial) logistic regression is a form of regression used when the dependent variable is a dichotomy and the independent variables are of any type, meaning they can take continuous forms (as with income) or categorical form. Logistic regression can be used to predict a dependent variable on the basis of a set of independents, to determine the per cent of variance in the dependent variable explained by the independents; or to rank the relative importance of independents. Most current research in the social care field has applied logistic regression models for the latter purpose. Another modelling technique to examine differences in characteristics between two groups is discriminant analysis. However, logistic regression has several advantages over discriminant analysis: it is more robust, as the independent variables do not have to be normally distributed, or have equal variance in each group; it does not assume a linear relationship between the independent and dependent variables; and a researcher may add (for example) explicit interactions and power terms. Unfortunately, the advantages of logistic regression come at a cost: it requires much more data to achieve stable, meaningful results in conjunction with the variability of the data. However, in the case of 'large' datasets this is not likely to be a problem, especially if the focus of the analysis is not on a very small subgroup of cases. Logistic, and other regression models may also be used to provide 'predictions' of outcomes for individuals in similar, but other, situations than where the model was developed.

Related to logistic regression is the multinomial logistic regression model, which handles the case of dependents with more than two classes. Hussein (2010b) used this model to investigate significant variation in social care workforce characteristics among different care settings. This method was more suitable than an ordered logit regression model due to the fact that care settings, such as domiciliary and day care, do not have a logical order by definition.

Other statistical models have been used in social care research, in some cases in an innovative manner; for example, McCann and colleagues' (2009) application of survival analysis, not only to examine mortality as traditionally used, but when investigating variations in care home residents' mortality rates. Survival analyses were also innovatively employed to gain further insight into patterns of variations in child protection practice in research conducted by Pugh and Jones (2004).

Multilevel modelling provides a powerful framework for exploring how average relationships vary across hierarchical structures, and policy makers are steadily beginning to recognise its value, particularly in relation to educational and geographical research. In the social care research field, multilevel modelling techniques were used by Hussein and colleagues (2009d) to identify the separate hierarchical effects of higher education institutions and social work students' personal characteristics on students' educational progression.

However, there is very limited use of other valuable techniques such as structural equation modelling (SEM). In a review of research that applied this method in the social care and social work field, Guo and colleagues (2009) identified 32 studies which used SEM, most of

them originating in the US. They further concluded that a primary use of SEM in social work is employing confirmatory factor analysis to assess the psychometric properties of instruments, with a small proportion involving testing structural relations in a full SEM or path analysis or testing latent growth curves. In the UK, there is little evidence of SEM application, although one piece of research in adult safeguarding has used exploratory factor analysis to examine types of mitigation claimed by social care staff referred to the POVA list (Hussein *et al*. 2009b).

DISCUSSION AND CONCLUSION

This review has provided some very useful examples of existing datasets which are currently or could potentially be used to enhance both the validity and coverage of research questions in the social care field. Research questions addressed by the studies examined here varied widely, covering topics including informal care provision, young offenders, child and adult protection, the supply of workers either through migration or new qualification, and the profile of and trends within the care workforce. Overall, utilising existing large datasets has the potential to bring great advantages to the social care research field with a minimal associated cost, although the costs of statistical expertise can be considerable. A great deal can be learned from existing diverse datasets enabling the building of a strong and social care-specific evidence base. Such evidence may also offer a springboard for more specific research questions in the social care field. Currently very limited options are considered by social care researchers in identifying, employing and maximising the benefits of existing datasets. There appear to be considerable missed opportunities in simply identifying and using existing datasets, whether these are national surveys or more localised administrative records, and in utilising in-depth statistical analyses. On a more positive note, during the past few years there appears to have been a growing interest in utilising such an approach in establishing relationships and exploring new research questions in the social care field.

Table 1 summaries the advantages of and important considerations in the use of existing large datasets as an approach in social care research. As discussed above, large datasets are already there waiting to be analysed; therefore the considerable costs associated with questionnaire design, data collection and entry are reduced. Existing datasets vary considerably, and they relate to different groups and cover a wide range of topics that are of interest to social care research. The volume of such data is expected to grow substantially in the coming years, with enhanced technology and the realisation by different administrative and governmental, social care and health bodies of the importance of accurate databases. Further opportunities exist in possible record linkage, as illustrated in some of the research summarised in this review.

Existing datasets offer information on large samples of people, or subjects, which may be difficult to achieve within the budget and timescale of most research projects. This is particularly useful when specific groups are considered, for example, older people from

Table 1 Advantages and considerations when employing the analysis of existing large datasets in social care research

Advantages	Readily available: reduces considerable costs
	Diverse datasets already exist, covering a variety of social care-related topics
	With technological advances more datasets are becoming available and record linkage may be facilitated
	Provides large samples, even for specific groups, which are difficult to attain in ordinary surveys
	Sample size allows more sophisticated statistical analyses to examine more precise research questions
	Provide accurate, valid and reliable evidence related to large groups of people
	Pose fewer ethical considerations than primary research
Considerations	Purpose and coverage of data may not match exact research questions
	Importance of developing a conceptual framework of analysis which maps to existing data
	Understanding the data may take some time
	Importance of data management and dealing with missing values
	May lack key variables or attributes
	Original sampling strategy needs to be considered
	Importance of choosing appropriate statistical techniques

BME groups or migrant workers, who may be difficult to access and will require considerable efforts to reach and achieve a good enough sample size. However, if a research question is specific to a very small subgroup of individuals, existing large datasets that did not deliberately sample this subgroup may include a very small number of them. In general, a large sample size offers a number of great advantages to any research, including the ability to consider more precise research questions and to use appropriate and sophisticated statistical techniques that will yield valid and reliable findings. Such methods also provoke fewer ethical issues, particularly when data anonymity and record-keeping have been considered from the outset of the research.

On the other hand, using existing large datasets poses a number of challenges and considerations. It is important that researchers realise the purpose, coverage and breadth of the data they use. They need to familiarise themselves with the data, how questions were asked, who provided the answers, and when they were asked; for example, whether people were asked about recent events or those that took place some time ago (see O'Keefe *et al*. 2007, who asked about mistreatment and neglect in the previous year). It is equally important to consider all available datasets which may be suitable and appropriate to the research questions and to weigh the value of each one against their limitations, as discussed in more detail earlier in this review.

Secondary data analysis still requires data management. Due to the fact that data collection has been designed and performed by parties other than the researchers, such work demands a great deal of attention in relation to the structure of the data, including coding and missing values. Coding may be of particular importance to administrative data, where data are usually collected for monitoring or observational purposes and may not include pre-coded values to the attributes that may be most valuable to the researchers. Thus, enough time should be allowed for activities such as re-coding, which will precede any actual data analysis.

The choice of analysis method is also important. This choice should be guided by a theoretical framework of analysis and by the initial phases of data exploration including sampling design, missing values analysis and distributions of variables of interest.

RECOMMENDATIONS

As highlighted by Sharland (2009), social care and social work research are at the crossroads of a number of intersections, including the individual and societal spheres, and interconnecting with a number of other systems including health, education and criminal justice. Utilising existing large datasets to examine a number of important research questions may enhance current knowledge and provide catalysts for new in-depth research questions. As this review illustrates, a number of useful large-scale databases exist, many of which have not been fully used within social care research when compared to nursing, demographic or population research from across the globe. Researchers have made a number of attempts to use such existing databases to examine new research

questions related to long-term care, child and adult protection, and the composition of the social care workforce, among others.

This review recommends that social care researchers take a closer look at what is available in terms of existing data which are relevant to their research questions. Further effort is required to identify, access and use such datasets including possible linkage of different data sources as well as constructing longitudinal analyses. Researchers need to make good use of the opportunities provided by existing large datasets and maximise their potential through exploring the depth and breadth of existing sources. Large datasets provide, by definition, information on large numbers of people or 'subjects'. If the focus of the research is on a very small subgroup of individuals, large datasets that are not designed to collect data from this specific group may not provide a large enough sample. With the employment of appropriate statistical techniques a stronger evidence base can be developed in relation to a variety of social care topics. This methods review has identified some excellent examples of the use of large datasets in the social care field. However, similar to Sharland's (2009) conclusions, it also found that these are very small in number and relatively uncommon within social care research. There is still great scope for using longitudinal data and combining different datasets, with a surprisingly significant under-use of social care services data (for exceptions, see Beadle-Brown *et al*. 2009 and Beecham *et al*. 2008).

It is also important to recognise that analysis of existing large datasets can be undertaken alongside other research methods, as has been demonstrated in a number of the studies mentioned above. Findings obtained from analysing large datasets can supplement and be contextualised through qualitative research. Multi-methods research design, in its wider and complementary format, is increasingly recommended. Using both secondary data analysis and other research approaches, including surveys and interviews for example, may provide richer insights into different aspects and theoretical concepts of research questions.

References

Andrew M (2005) Social capital, health, and care home residence among older adults: a secondary analysis of the Health Survey for England 2000, *European Journal of Ageing*, 2, 2, 137–148.

Beadle-Brown J, Mansell J, Knapp M, Beecham J (2009) Residential services in Europe – findings from the DECLOC study, *International Journal of Integrated Care*, 9, e9.

Beecham J, Knapp M, Fernández J, Huxley P, Mangalore R, McCrone P, Snell T, Winter B, Wittenberg R (2008) *Age Discrimination in Mental Health Services*, PSSRU Discussion Paper 2538, Personal Social Services Research Unit, London School of Economics and Political Science, London.

Berger A, Berger C (2004) Data mining as a tool for research and knowledge development in nursing, *Computer Informatics in Nursing*, 22, 123–131.

Cangiano A, Shutes I, Spencer S, Leeson G (2009) *Migrant Care Workers in Ageing Societies: Research Findings in the United Kingdom*, COMPAS, Oxford.

Carmichael F, Charles S, Hulme C (2010) Who will care? Employment participation and willingness to supply informal care, *Journal of Health Economics*, 29, 182–190.

Cooper C, Regan C, Tandy A, Johnson S, Livingston G (2007) Acute mental health care for older people by crisis resolution team in England, *International Journal of Geriatric Psychiatry*, 22, 263–265.

Curtis L, Netten A (2005) *Unit Cost of Health and Social Care*, Personal Social Services Research Unit, University of Kent, Canterbury, Kent.

Dahlberg L, Demack S, Bambra C (2007) Age and gender of informal carers: a population-based study in the UK, *Health and Social Care in the Community*, 15, 5, 439–445.

Del Bono E, Sala E, Hancock R (2009) Older carers in the UK: are there really gender differences? New analysis of the Individual Sample of Anonymised Records from the 2001 UK Census, *Health and Social Care in the Community*, 17, 3, 267–273.

Dickens J, Howell D, Thoburn J, Schofield G (2007) Children starting to be looked after by local authorities in England: an analysis of inter-authority variation and case-centred decision making, *British Journal of Social Work*, 37, 597–617.

Dobson J, Salt J (2009) Foreign recruitment in health and social care: recent experience reviewed, *International Journal of Migration, Health and Social Care*, 2, 3–4, 1747–9894.

Doran T, Drever F, Whitehead M (2003) Health of young and elderly informal carers: analysis of UK census data, *British Medical Journal*, 327, 1388.

Eborall C, Griffiths D (2008) *The State of the Adult Social Care Workforce in England 2008. The Third Report of Skills for Care's Skills Research and Intelligence Unit*, Skills for Care, Leeds.

Evaluation of the New Social Work Degree Qualification in England Team (2008). *Findings from the Evaluation of the New Social Work Degree Qualification: Research Summary I*, Social Care Workforce Research Unit, King's College London, London.

Felce D, Perry J, Romeo R, Robertson J, Meek A, Emerson E, Knapp M (2008) Outcomes and costs of community living semi-independent living and fully staffed group homes, *American Journal on Mental Retardation*, 113, 2, 87–101.

Fox D, Holder J, Netten A (2010) *Personal Social Services Survey of Adult Carers in England – 2009–10: Survey Development Project*, Technical report, Personal Social Services Research Unit, University of Kent at Canterbury.

Genovese CR, Lazar NA, Nichols T (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate, *Neuroimage*, 15, 4, 870–878.

Gill P, Shankar A, Quitke T, Freemantle N (2009) Access to interpreting services in England: secondary analysis of national data, *BMC Public Health*, 9, 12, doi:10.1186/1471-2458-9-12.

Guo B, Perron B, Gillespie D (2009) A systematic review of structural equation modeling in social work research, *British Journal of Social Work*, 39, 1556–1574.

Hassiotis A, Ukoumunne O, Byford S, Tyrer P, Harvey K, Piachaud J, Gilvarry K, Fraser J (2001) Intellectual functioning and outcome of patients with severe psychotic illness randomised to intensive case management: Report from the UK700 trial, *British Journal of Psychiatry*, 178, 166–171.

House of Commons (2009) Official Statistics: 2011 Census Questions. Public Administration Select Committee. http://www.gscc.org.uk/NR/rdonlyres/B5A5B087-E7B9-471C-BAAF-207DA1FBE1DA/0/Progression_analysis_FT_UG.pdf.

Hussein S (2010a) Modelling pay in adult care using linear mixed-effects models, *Social Care Workforce Periodical*, Issue 7, June 2010; web published at http://www.kcl.ac.uk/sspp/kpi/scwru/pubs/periodical/issues/scwp7.pdf.

Hussein S (2010b) Adult day care workforce in England. *Social Care Workforce Periodical*, Issue 4, February 2010; published online at http://www.kcl.ac.uk/scwru/pubs/periodical/.

Hussein S (2011) Migrant workers in long term care: evidence from England on trends, pay and profile, *Social Care Workforce Periodical*, Issue 12, March 2011; published online at http://www.kcl.ac.uk/sspp/kpi/scwru/pubs/periodical/issues/scwp12.pdf.

Hussein S, Moriarty J, Manthorpe J, Huxley P (2008) Diversity and progression among students starting social work qualifying programmes in England between 1995 to 1998: a quantitative study, *British Journal of Social Work*, 38, 8: 1588–1609.

Hussein S, Stevens M, Manthorpe J, Rapaport J, Martineau S, Harris J (2009a) Banned from working in social care: a secondary analysis of staff characteristics and reasons for their referrals to the POVA list in England and Wales, *Health & Social Care in the Community*, 17, 5, 423–433.

Hussein S, Martineau S, Stevens M, Manthorpe J, Rapaport J, Harris J (2009b) Accusations of misconduct among staff working with vulnerable adults in England and Wales: their claims of mitigation to the barring authority, *Journal of Social Welfare & Family Law*, 31, 1, 17–32.

Hussein S, Manthorpe J, Stevens M, Rapaport J, Harris J, Martineau S (2009c) Articulating the improvement of care standards: the operation of a barring and vetting scheme in social care, *Journal of Social Policy*, 38, 2, 259–275.

Hussein S, Moriarty J, Manthorpe J (2009d) *Variations in Progression of Social Work Students in England: Using Student Data to Help Promote Achievement: Undergraduates Fulltime Students' Progression on the Social Work Degree*, General Social Care Council, London.

Hussein S, Stevens M, Manthorpe J, Rapaport J, Martineau S, Harris J (2010a) Using governmental data records for research: a case study of understanding characteristics and reasons for social care workers' referrals to the POVA List in England and Wales, *Radical Statistics*, 100, 11–16.

Hussein S, Stevens M, Manthorpe J (2010b) *International Social Care Workers in England: Profile, Motivations, experiences and Future Expectations*, February 2010, Final Report to the Department of Health, Social Care Workforce Research Unit, King's College London, London.

Huxley P, Evans S, Burns T, Fahy T, Green J (2001) Quality of life outcome in a randomized controlled trial of case management, *Social Psychiatry and Psychiatric Epidemiology*, 36, 5, 249–255.

ID Research (2002) *Gay and Lesbian Census*, ID Research, London.

Kavanagh S, Knapp M (2002) Costs and cognitive disability: modelling the underlying associations, *British Journal of Psychiatry*, 180, 120–125.

Knapp M, Romeo R, Beecham J (2009) Economic cost of autism in the UK, *Autism*, 13, 3, 317–336.

Malley J, Netten A (2008). Measuring and monitoring outputs in social care: the problem of measuring quality, *International Review of Administrative Sciences*, 74, 4, 535–553.

McCann M, O'Reilly D, Cardwell C (2009) A Census-based longitudinal study of variations in survival amongst residents of nursing and residential homes in Northern Ireland, *Age and Ageing*, 38, 711–717.

Morgan C, Ahmed Z, Kerr M (2000) Health care provision for people with a learning disability: Record-linkage study of epidemiology and factors contributing to hospital care uptake, *British Journal of Psychiatry*, 176, 37–41.

Mulholland J, Anionwu E, Atkins R, Tappern M, Franks P (2008) Diversity, attrition and transition into nursing, *Journal of Advanced Nursing*, 64, 1, 49–59.

Netuveli G, Wiggins R, Hildon Z, Montgomery S, Blane D (2006) Quality of life at older ages: evidence from the English longitudinal study of aging (wave 1), *Journal of Epidemiology and Community Health*, 60, 357–363.

O'Keeffe M, Hills A, Doyle M, McCreadie C, Scholes S, Constantine R, Tinker A, Manthorpe J, Biggs S, Erens B (2007) *UK Study of Abuse and Neglect of Older People Prevalence Survey Report*. Department of Health, http://www.dh.gov.uk/en/Publicationsandstatistics/%20Publications/PublicationsPolicyAndGuidance/DH_076197.

Office for National Statistics (2006) *Sexual Orientation and the 2011 Census*, Office for National Statistics, London.

Pollack CD (1999) Methodological considerations with secondary analysis, *Outcomes Management for Nursing Practice*, 3, 147–152.

Price E (2011) *LGBT Sexualities in Social Care Research, SSCR Methods Review 2*, NIHR School for Social Care Research, London.

Pugh R, Jones P (2004) Survival Analysis in Social Work Research, *British Journal of Social Work*, 34, 6, 907–914.

Purdam K, Wilson A, Afkhami R, Olden W (2008) Surveying sexual orientation: Asking difficult questions and providing useful answers, *Culture, Health and Sexuality*, 10, 2, 127–141.

Robertson J, Emerson E, Hatton C, Elliott J, McIntosh B, Swift P, Krijnen-Kemp E, Towers C, Romeo R, Knapp MRJ, Sanderson H, Routledge M, Oakes P, Joyce T (2005) *The Impact of Person Centred Planning*, Department of Health, London.

Shah A (2009) The 'Count Me In' psychiatric in-patient census for 2007 and the elderly: evidence of improvement or cause for concern?, *Psychiatric Bulletin*, 33, 201–203.

Sharland E (2009) *Strategic Adviser for Social Work and Social Care Research*, Report to the Economic and Social Research Council Training and Development Board.

Shivram R, Bankart J, Meltzer H, Ford T, Vostanis P, Goodman R (2009) Service utilization by children with conduct disorders: findings from the 2004 Great Britain child mental health survey, *European Child and Adolescent Psychiatry*, 18, 9, 555–563.

Sin C (2006) The feasibility of using national surveys to derive samples of older people from different ethnic groups in Britain: Lessons from 'piggy-backing' on the Family Resources Survey, *International Journal of Social Research Methodology*, 9, 1, 15–28.

Sondhi A, Huggins R (2005) Towards an effective social care model for arrest referral: Implications for criminal justice interventions for problem drug users, *Drugs: Education, Prevention and Policy*, 12, 3, 189–195.

The NHS Information Centre (2011) *Personal Social Services Adult Social Care Survey: Guidance Document – 2011–12*, The Health and Social Care Information Centre, London.

Tyrer P, Oliver-Africano P, Ahmed Z (2008) Risperidone, haloperidol and placebo in the treatment of aggressive challenging behaviour in patients with intellectual disability: a randomised controlled trial, *Lancet*, 371, 57–63.

UK700 Group (2000) Cost-effectiveness of intensive v. standard case management for severe psychotic illness: UK700 case management trial, *British Journal of Psychiatry*, 176, 537–543.

Vostanis P, Bassi G, Meltzer H, Ford T, Goodman R (2008) Service use by looked after children with behavioural problems: Findings from the England survey, *Adoption & Fostering*, 32, 3, 23–32.

Young H, Grundy E (2008) Longitudinal perspectives on caregiving, employment history and marital status in midlife in England and Wales, *Health and Social Care in the Community*, 16, 4, 388–399.