

Encouraging the wider and more creative use of administrative data in the UK - the 'Administrative Data Liaison Service'

Chris Dibben¹, Heather Gowans², Mark Elliot³, Chelsie Anttila⁴, Paul Boyle⁵, Jane Kaye⁶, David McLennan⁷, Michael Noble⁸, George Smith⁹, Kate Wilkinson¹⁰

¹Department of Geography and Geosciences, University of St Andrews, e-mail: cjld@st-andrews.ac.uk

²ETHOX Centre, Division of Public Health, University of Oxford, e-mail: heather.gowans@dphpc.ox.ac.uk

³Cathie Marsh Centre for Census and Survey Research, Social Sciences Department, University of Manchester, email: Mark.J.Elliot@manchester.ac.uk

³Social Disadvantage Research Centre, Department of Social Policy and Social Work, University of Oxford, email: chelsie.anttila@socres.ox.ac.uk

⁴Department of Geography and Geosciences, University of St Andrews, email: p.boyle@st-andrews.ac.uk

⁵ETHOX Centre, Division of Public Health, University of Oxford, email: jane.kaye@ethox.ox.ac.uk

⁶Social Disadvantage Research Centre, Department of Social Policy and Social Work, University of Oxford, email: david.mclennan@socres.ox.ac.uk

⁷Social Disadvantage Research Centre, Department of Social Policy and Social Work, University of Oxford, email: Michael.Noble@socres.ox.ac.uk

⁸Social Disadvantage Research Centre, Department of Social Policy and Social Work, University of Oxford, email: George.Smith@socres.ox.ac.uk

⁹Social Disadvantage Research Centre, Department of Social Policy and Social Work, University of Oxford, email: Kate.Wilkinson@socres.ox.ac.uk

Abstract

Administrative data has the potential to provide a relatively cheap, potentially less intrusive and yet comprehensive resource for research in the UK. Many European countries, notably those of the north, have placed a much greater emphasis on the development of this type of data than the UK and as a consequence have managed to replace national censuses and major surveys, reducing costs to the taxpayer and the burden on citizens from requests for information. Although the use of this type of data has increased in the UK, it still lags behind European neighbours. Recent 'losses' of data by UK government departments and sensitivities amongst the UK population to the concept of central registers of citizens and the associated policy of identity cards have made this a particularly sensitive issue.

In an attempt to address this problem, the UK Economic and Social Research Council (ESRC - the UK's national economic social science funding body) has funded a consortium of Universities to try and negotiate a pathway through the various impediments to the wider use of administrative data. The Administrative Data Liaison Service (ADLS) is provided by the Universities of St Andrews, Oxford and Manchester, and will act as a focal point for knowledge about the availability of administrative data, their suitability for specific research purposes and the procedures required to gain access to and use such data. The service is made up of experts in the legal and ethical use of 'personal information', data disclosure risk and research uses of administrative data. It will work in tandem with government departments and agencies, seeking to develop and improve the use of administrative data resources for research purposes within a clear and transparent ethical and legal framework.

Key words

Administrative data, ethics, data disclosure, anonymity

1. Introduction

Administrative data (that is, information collected primarily for administrative purposes) has long contributed to central government and other statistics in the UK. In the last 10-15 years, technological advancement has facilitated the formation of very large administrative databases held by central and local government, and by specialist agencies across the UK. The existence of such databases raises the possibility that administrative data could become a core resource for academic research.

Many northern European countries (notably Denmark, Netherlands, Sweden and Finland) are further ahead in this regard, as they have placed much greater emphasis on developing national 'register data' from the 1980s onwards, to replace national censuses and major surveys. For example, *Statistics Denmark* now bases most of its national statistics on 'register data' which can be linked both longitudinally and between registers of different types (e.g. health, education, income). Controlled access to these population-

wide, longitudinal registers *at an individual level* have allowed studies of the precursors of events (for example Christofferson *et al's* 2003 study on attempted suicide among young people). In addition, Danish surveys are frequently supplemented with register data on income, health, welfare benefits, housing etc. allowing objective information to be compared to the responses from survey members. These extremely valuable resource are as yet unavailable to UK academics on a routine basis (Boyle *et al.* 2004a,b).

Datasets with very similar potential are now beginning to emerge in the UK. Many Government departments manage large-scale databases, however these are largely department or function specific although some progress has been made in linking these together (e.g. the Department for Work and Pension (DWP) have linked welfare benefits and tax data). Such developments clearly lay the foundation for the Office for National Statistics proposals (ONS, 2003) to use such administrative data to supplement, or even replace, the decennial census after 2011, and they represent a rich potential resource for the UK social science research community. In the UK there appears to be growing cross government support for data-sharing between public agencies with appropriate safeguards to protect privacy (Department for Health, 2007). A Cabinet Office Committee on data-sharing has been established, chaired by the Minister for the Cabinet Office and for Social Exclusion, and a recent Council for Science and Technology report highlighted the potential benefits of such sharing (Council for Science and Technology, 2005). The opportunities to create extensive, powerful datasets based on administrative data are therefore growing.

However, academic social science in the UK has so far been at the margins of these developments. The reasons for this were explored in a recent audit (Jones and Elias, 2006) that found that *'knowledge about the suitability of different administrative data resources for various research purposes is limited and fragmented'* (p.81). In particular, knowledge of: the data available and its potential; the procedures for accessing it; methods for exploiting it; and its quality and suitability needed to be harnessed and coordinated if the UK social science community was to fully realise the potential of these powerful new research tools. The ESRC have been keen to place academic research more centrally within this process. As a result they have funded a number of research programmes and one of these is the Administrative Data Liaison Service (ADLS) which is discussed in this paper. We first describe the context within which the ADLS must function. We then outline the ADLS itself.

2. Administrative data within the UK

UK government departments are developing and maintaining major databases that have the potential to provide important indicators and datasets for social science research. Education, labour market, health, business, and demographic research could clearly be informed by data held in these welfare, tax, health and educational record systems. However these databases are largely department or function specific (not, as yet, cross-department or function) and have, on the whole, not been linked across these functional areas and therefore their potential, as a social science research tool, as yet not fully exploited.

To date, research using administrative datasets in the UK has tended to be limited to that supported by specific research commissions by government departments themselves. This means that methods for working with these types of data and procedures to handling them are relatively underdeveloped. Analysis has also been largely restricted to questions set by funders. Two areas of research, education research using the school level results (the National Pupil Database) and the health research using information on hospital admissions (the Hospital Episodes Statistics), have atypically seen more extensive use of administrative data, however for the other large administrative datasets, there have only been a small number of individual academic research groups working with them. The Social Disadvantage Research Centre (SDRC) at the University of Oxford is one example. They pioneered the use of such administrative data in the late 1990s, working with individual level datasets including welfare benefits, tax credit data, pupil level educational data hospital records etc. They also created wholly new datasets (e.g. the first ever national database of recorded crime for the whole of England, based on individual crime records from all 39 police forces) and linked these longitudinally (at individual level) to examine trends (e.g. Noble, Evans, Dibben and Smith 2001; Evans, Noble, Wright, Smith, Dibben, 2002). However this type of work has involved negotiations with government departments and other agencies across the UK, and the development of 'data contracts' with data suppliers to hold and use such data. It is therefore not a route that many research groups have taken. As a result, for many administrative datasets, there are only small clusters of researchers who actively used them. The knowledge of how to access them and skills to use them, are therefore limited to small groups and their use by the wider research community is very limited.

3. Public concern in the UK

In the background for all further development of administrative data use in research, lies the not atypical concern amongst the UK population about the use of personal information. This is, of course, a common concern across Europe (Thomas and Walport, 2008), however it has been exacerbated in the UK by a series of accidental loss of personal data by government departments, in particular the theft of a Ministry of Defence (MoD) laptop containing information on armed forces personnel in 2008 and the 'loss' of Discs containing Child Benefit Records (a universal benefit paid to the parents of children under 16 in the UK) in 2007 by HM Revenues and Customs. Government has responded to these events with a series of reviews (Burton Review into the MoD laptop loss and Poynter Review into the loss of Child Benefit records), however it remains to be seen whether these reviews, and the improved data handling and sharing practises resulting, will increase the public's confidence in the ability of government to safely and acceptably handle personal data.

At the same time as this scrutiny of the government's ability to safely handle personnel data, there has also been a continuing discourse, both public and in the media, that centres on the extent to which administrative data is being used by the state and other agencies to scrutinise the public. Often coming under the banner of the fear of a 'surveillance society', these discourses focus on concerns over the extent to which the use of administrative data may be impacting individual freedoms.

These events and public discourse have generated an environment that is not especially conducive to arguments for the extended use of administrative data for research purposes. On the whole they have tended to lead to an environment where the risks associated with the extension of these types of uses are very salient but the potential benefits are not.

4. The Legal context for administrative data exploitation in the UK

The legal context for the sharing of data for research processes in the UK is a complex issue. There is no single source of law and as a result the legal basis for such sharing may not be clear cut (Thomas and Walport, 2008). The laws regulating the collection, use and sharing of personal data in the UK are governed by the common law, European legislation and UK statutes, and this legal basis is important in relation to the research uses of administrative data in the UK.

Briefly stated, the common law recognizes that a duty of confidentiality arises in certain relationships (for example, in the doctor and patient relationship), and personal information cannot normally be disclosed to a third party without the consent of the data subject. In respect of public authorities, administrative law governs their actions, and a disclosure of information by a public authority may be prevented by lack of function permitting the authority to do so.

The European Union Data Protection Directive relates to the protection of privacy and to the processing of all personal data collected for, or about EU citizens, and it was implemented in the UK in the form of the Data Protection Act 1998 (DPA). The DPA regulates the collection, use, distribution, retention and destruction of personal data in the UK. It also establishes various rights for individuals, including a right of access to personal information held about them. However, when personal information is further processed purely for research purposes, the data is exempt from the data subject's right of access and can be held indefinitely, if it is processed in compliance with the relevant conditions, and the results of the research or statistics are not in a form which causes harm to the data subject(s).

The Human Rights Act 1998 (HRA) incorporates the European Convention on Human Rights, and personal data is protected by Article 8 of the Convention as part of an individual's right to respect for private life. Any infringement of privacy must be necessary, in accordance with the law and proportionate. UK law must provide sufficient safeguards to ensure that data sharing does not infringe the HRA, and that there is a legal basis for any infringement.

More recent legislation in the form of the Statistics and Registration Service Act 2007 (SRS) is mainly concerned with the structure of the organisations that will deliver statistics and registration services, and it introduced a Statistics Board responsible for promoting and safeguarding the production, publication and quality of official statistics in the UK. The Statistics Board therefore has power to promote statistical research by providing access to data held by it where it is lawful to do so. The SRS, however, provides that personal information which is held or disclosed by the Statistics Board is

confidential, and any disclosure of the same is unlawful. One exception to this, is where there is disclosure of personal information to ‘an approved researcher’ to whom the Statistics Board has granted access for the purpose of statistical research. ‘Approved researcher’ has not been defined in the legislation.

Despite various types of legislation, however, uncertainties about the information laws in the UK do arise, and can therefore pose concerns over data sharing for research purposes in the UK. One example of where problems have arisen is on the application of the DPA to the use and disclosure of personal health data, specifically in respect of the consent required before data may be lawfully shared with another. Although the DPA makes provision for some circumstances where it may be necessary to process an individual’s health data without consent (for example, in some types of research where it would be impractical to obtain consent for the use of old health records), it is the first data protection principle of the DPA that data must be processed ‘lawfully’ (that is, in compliance with the common law on confidentiality, which binds all medical practitioners not to release personal medical data). In this context, the Information Commissioner has therefore issued legal guidance on the use of confidential medical data.

In the UK, in order to avoid confusion and uncertainty about the law, anonymised data should be used wherever possible so as to ensure compliance with the law. This however can be problematic because full anonymisation of data is often impractical where, for example, identifiable data is needed for linkage between data sets; when identifiers contain information that is useful to the researcher; or when it is too laborious to anonymise the data set. For many research applications, where it is useful to have a series of indicators for individuals, for different points in time and for specific geographical areas; an individual may be recognisable even in a dataset ‘anonymised’ through the removal of names and addresses. The resulting risk of disclosure is therefore a key issue for the potential sharing of administrative data.

5. Anonymisation and Disclosure

One of the issues that the stewards of administrative data must face when they are considering dissemination for research purposes is the problem of statistical disclosure. This problem has been studied extensively with respect of official statistics (see e.g. Elliot 2004) and might be roughly defined as the disclosure of information about a specific population unit from anonymised data.

Controlling disclosure is a complex and inexact process and an axiom of the field is that it is not possible to reduce the risk of disclosure to zero. There are many different disclosure control mechanisms, but these can be broadly categorised into *restrictions of access* and *restrictions of the data*. Data restrictions include releasing samples rather than population data, reductions in the level of detail and perturbation. Access restrictions include only allowing access to particular licensed researchers and only allowing access through a particular software system.

The other side of the disclosure coin is *data utility*. All methods of disclosure control reduce utility in some way and therefore sophisticated disclosure risk analysis attempts to optimise the residual utility whilst minimising the risk. One of the functions of ADLS will be to provide stewards of administrative data with guidance on the appropriate mix and levels of disclosure control for their data.

6. The ADLS service

In an attempt to address some of the issues and problems outlined above, the UK's ESRC has funded a consortium of Universities to try and negotiate a pathway through the various impediments to the wider use of administrative data. The Administrative Data Liaison Service will be provided by the Universities of St Andrews, Oxford and Manchester. It is design to comprise three components: service delivery; service development; and dissemination and engagement. This will be achieved through three strands of work:

- A. **The core service** to the academic community and administrative data custodians.
- B. **A service development function:** the development of administrative data as a research resource.
- C. An active **dissemination and engagement** strategy.

The core service

A 'point of contact' service to the academic community and government departments will be provided through a **core service unit** based in the University of St Andrews. The staff in the core service unit will have day-to-day contact with members of the academic community interested in accessing data and with administrative data custodians. The staff will be available for advice on the requirements for holding data (e.g. data security, IT architecture etc.). They will be able to provide the relevant data dictionaries and protocols for applying for access to the data. They will know when the protocols need to be submitted and will be able to offer advice on how to complete them. They will also be in frequent contact with the relevant government departments ensuring that departments need to communicate with one central unit on social science access to administrative data and that the ADLS is up-to-date with developments concerning administrative data.

Direct communication with the core unit will be via email and telephone, though as the service develops, an increasing amount of information will be made available through the ADLS web site, including a compilation of existing resources as well as those developed by the consortium (including: protocols, guidance documents, frequently asked questions, data dictionaries and exempla case studies). The long-term aim will be to maximise the reliance on the website material and to use direct communication less and less, as the shared professional knowledge increases.

In order to help individuals understand the legal and ethical landscape affecting administrative data use, guidance will be provided on the website. This information will be distilled and organised into a cascade form of drop-down menus, 'user -friendly', intuitive, and informative for a wide audience of stakeholders. The team will ensure

that the website is kept up-to-date with changes in the legal and regulatory requirements and with the impact of technological change on the use of administrative data. It will also identify the strengths and shortcomings of the existing legal mechanisms, rules and frameworks by documenting researchers' queries. The ADLS website will also contain information on the available administrative datasets, their relative weaknesses and strengths and provide broad guidance on their use. The materials for this part of the service will be developed as part of the second work component.

Service development: The broader development of administrative data as a research resource

Although administrative data has enormous potential as a research tool, it is not always in a form suitable for research. There are questions for data custodians about whether their data can be used for research and whether they can be made available to *bona fide* academic researchers. The second strand of work will involve close collaboration with government departments to extend the potential of such data.

This work will have two purposes:

1. To increase the administrative data knowledge base. This will involve the production of guidance documents that will be available through the core service.
2. To advance the development of administrative data more generally.

A major focus of this work will be the exploration of significant administrative data research issues. This work will include:

1. Testing the strengths and weaknesses of existing datasets as research resources; producing quality reviews of administrative data sources and developing methodologies for addressing their weaknesses.
2. Examining the possibilities of data linkage to produce more powerful combined datasets.
3. Assessing the particular disclosure risks associated with administrative data use and establishing methods for ameliorating those risks (policies, procedures, use models).
4. Legal Analysis of administrative data dissemination and use.

Until now there has probably been an unhelpful focus, in the UK, on the difficulties of using administrative for research. This strand of work will look realistically at potential biases and questions of validity arising from the use of such data. It will assess the main individual level datasets, assessing and reviewing their quality. The findings from this work will then be published on the service website. Methodologies to address some of these research shortcomings will also be initiated. This will particularly focus on the administrative records as proxy indicators for 'useful' research variables and how methods of modelling and estimation can be used to improve them. For example, how various benefits data can be combined to produce a reliable estimate of low income households or what hospital admissions can be used to produce estimates of the probability that an individual abuses alcohol?

As discussed, although administrative datasets in the UK have been linked in powerful ways within departments, this linkage has rarely been extended across departments. This means that in the administrative data has not played the significant role in research that it plays in some Scandinavian and Northern European countries where, in affect, it replaces surveys and even censuses. This second strand of work will explore the legal, methodological and research possibilities of a wider linkage across departments. This consortium has considerable experience of such negotiations and, in particular, the linkage of health and socio-economic datasets to produce powerful datasets capable of exploring health inequalities etc.

Minimising the risk of disclosure from administrative data is of great importance if the public is to retain confidence in its use for research. The relative risks of breaches of confidentiality will be explored by analysing various aspects of an increased access to administrative data for the academic community. Guidelines and policies on confidentiality and disclosure risk and control for data custodians who are making their data available will be developed. Some disclosure risk assessments of exemplar datasets will be conducted as well as scenario analyses.

The legal landscape that confronts academic researchers using administrative data is complex and continually evolving. The service will identify, collate, and critically examine the existing framework of laws and regulations applicable to the use of administrative data in UK, as well as other forms of guidance and secondary sources. It will also document in detail, the legal obligations and responsibilities of all stakeholders (which includes researchers as well as regulatory bodies) and the procedures required by law for the use of administrative and statistical information in research (for example, the SRS Act 2007). Finally, it will feed these findings to individuals through the website and advisory service, as well as appropriate organisations in order to ensure that social science research using administrative data is adequately supported in law and through appropriate regulatory frameworks and mechanisms

This work will be carefully documented and the findings published in the form of guides on the ADLS website. The programming and syntax used in these exploratory exercises as well as the newly created variables will be made available to future users where appropriate and, where possible, departments will be encouraged to archive derived datasets so that they may be used by other groups of researchers.

Dissemination and engagement

There is a variable knowledge base within higher education and other research organisations with a minority of highly skilled and experienced users of administrative data, with a much larger group that have very little experience and some scepticism towards this research resource. The level of skills and knowledge does vary between disciplines. So, for example, education researchers in the UK, are often very familiar with administrative data emanating from schools, while other social scientists are quite likely to have never used any administrative data in their work. The third strand of work will involve disseminating information on and engaging academic researchers with the potential uses of administrative data.

This will be achieved through: Biennial workshops held throughout the UK on the use of administrative data (covering ethical, legal, data security and administrative data research), Working papers – outlining findings from component, Conference and journal papers (on the ADLS itself as well as ethical, legal, data security and administrative data research) and the promotion of academic research, based on administrative data, within government.

7. Conclusion

Administrative data has the potential to provide the research community with a relatively cheap, potentially less intrusive and yet comprehensive research resource. The potential research value is difficult to estimate but is undoubtedly significant. However, it is not an easy resource to use and one that also requires careful consideration of its quality and reliability and the ethical and legal issues that underpin its use. There is also, at the moment, uncertainty over the extent to which researchers will be given access to administrative datasets. The context for the new ADLS is therefore challenging but the rewards to the UK research community potentially large.

References

- Boyle PJ, Cullis A, Flowerdew R and Gayle V 2004a UK Data Audit Phase I, Report to the ESRC Research Resources Board
- Boyle PJ, Cullis A, Flowerdew R and Gayle V 2004b UK Data Audit Phase II, Report to the ESRC Research Resources Board
- Christoffersen, M.N., Poulsen, H.D., Nielsen, A., (2003) Attempted suicide among young people, *Acta Psychiatrica Scand* 108: 1-9.
- Council for Science and Technology (2005) Better use of Personal Information: opportunities and risks Council for Science and Technology.
- Department for Health (2007) Informing Healthier Choices: Information and Intelligence for Healthy Populations Department for Health.
- Elliot, M. J. (2004). 'Statistical Disclosure Control', In *Encyclopaedia of Social Measurement*. New York: Elsevier.
- Evans, M., Noble, M., Wright, G., Smith, G.A.N., Lloyd, M. & Dibben, C., (2002) *Growing Together or Growing Apart? Geographic Patterns of Change in IS and JSA-IB Claimants in England 1995-2000* (The Policy Press).
- Jones P, and Elias P. (2006) Administrative data as research resources: a selected audit Economic And Social Research Council.
- Noble, M., Evans, M., Dibben, C. and Smith, G.A.N., (2001) *Changing Fortunes: Geographic patterns of Income Deprivation in the late 1990s*. Department of the Environment, Transport and the Regions, Regeneration Series.
- Office for National Statistics, (2003) 'Proposals for an Integrated Population Statistics System', Office for National Statistics Discussion Paper, October 2003.
- Thomas and Walport (2008) *Data Sharing Review Report*, Ministry of Justice.