Enhancing longitudinal surveys by linking to administrative data
Lisa Calderwood, Centre for Longitudinal Studies and Carli Lessof, National Centre for Social Research

1. Introduction

Linking survey data to administrative data has become increasingly common in the UK and elsewhere. This growth has been encouraged by the expansion of what is available and technically possible. At the same time, there has been a rise in concern about linkages, and an increase in the ethical and regulatory constraints that surround the process.

The primary motivation in linking to administrative records has been to enhance the survey data in order to provide greater opportunities for research. Administrative data can be used to supplement and/or validate survey data. However, in order to fully exploit such opportunities there are substantive issues about how survey and administrative data should be used together. There are also practical issues that need to be considered when linking survey data to administrative data. These are both technical issues regarding how the data should be linked and ethical and legal issues about consents and disclosure risk.

Many of the potential uses of administrative data and practical issues that are associated with linking to administrative data apply to any type of survey. However, there are also particular considerations that are unique to longitudinal surveys or different in the case of longitudinal surveys.

This chapter will also argue that as well as enhancing the survey data, administrative data can also be used to enhance the survey quality in general, for example, by helping to keep track of respondents. Several possible uses of administrative data for improving survey quality will be considered and it will be argued that such methods are particularly beneficial for longitudinal surveys.

This paper begins by discussing the different types of linkage of administrative with survey data (section 2) and summarising existing linkages with longitudinal studies in the UK (section 3).

The main focus of the paper is on discussing the different motivations for linking survey data to administrative data and how these differ for longitudinal surveys compared with cross-sectional surveys (section 4).

The practical issues involved in linkage are discussed, again with reference to differences between longitudinal and cross-sectional surveys (section 5).

The final section (section 5) summarises and concludes.
2. **Combining survey and administrative data**

Administrative data is information that is routinely collected by organisations, institutions, companies and other agencies in order that the organisation can carry out, monitor, archive or evaluate the function or service it provides.

As with survey data, administrative data may be either longitudinal or cross-sectional in nature. Many administrative datasets collect store information by spell e.g. period of welfare benefit receipt or time spent in hospital. Such datasets are inherently longitudinal as successive spells for a given individual can be linked with each other so that change can be observed over time.

The focus of this paper is on linking longitudinal surveys to either cross-sectional or longitudinal administrative data. However, it is also possible to create a longitudinal data resource by linking a cross-sectional survey to administrative data - either to cross-sectional administrative data collected at a different time point than the survey data or to longitudinal administrative data. Longitudinal data resources can also be created by linking different administrative datasets to each other or linking census data to longitudinal administrative data. Examples of these different combinations of linkage are discussed in the next section.

3. **Administrative data and survey linkages in the UK**

3.1 **Administrative data in the UK**

Administrative data held by government departments and public sector organisations are among those most often used by surveys in the UK in linked studies. They hold a range of information about varied and fundamental aspects of people’s lives. This includes key demographic information such as birth and death, data about our experience of illness and our receipt of health services, participation in and outcomes from the education system, as well as employment spells, income earned, benefits received, tax paid and National Insurance contributions made. The exploitation of these data for research purposes has been greatly facilitated by the increasing computerisation of administrative records and by technological advances, which allow greater volumes of data to be stored and processed.

Some of the most common examples are set out in Figure 1.

---

**Figure 1:** Examples of administrative data sets that are linked to survey data in the UK

Then National Health Service Central Register (NHS-CR) is a central record of all NHS patients registered in England and Wales. Its main purpose is to ensure that medical records are passed between Health Authorities effectively and that General Practitioners are paid correctly. The data set includes name and surname, date and place of birth, sex, date of death, age of deceased, cause of death, occupation of deceased and the health authority to which a
A person is registered. Medical information is not held on the NHSCR. Linkage to NHSCR provides key information necessary for survival analysis.

All the UK Cancer Registries collect information each year on every new diagnosis of cancer occurring in their populations. They collate data on individual patients from multiple sources and over long time periods. These sources include district general hospitals, cancer centres, hospices, private hospitals, cancer screening programmes, other cancer registers, primary care, nursing homes and death certificates. Data are frequently collected from several sources within an institution, for example pathology departments, medical records and radiotherapy databases. Cancer registration in England is conducted by nine regional registries, which collect and collate data on cancers resident in their area, and submit a standard dataset on these registrations to ONS. The national cancer intelligence centre (NCIC) at ONS coordinates the national collation of cancer registration data and carries out analysis and research.

The Hospital Episodes Statistics (HES) register contains an entry for each hospital episode experienced by any patient using the National Health Service in England since 1989. A hospital episode is defined as any hospital admission as an inpatient. The register holds data for each episode such as the admission date, reason for admission, nature of illness, maternity care, waiting time, treatment given and discharge date. The Information and Statistics Division of NHS Scotland holds similar information on patient care delivered by NHS hospitals in Scotland, such as reason for the visit, length of stay and waiting time. Both in-patient and out-patient visits to hospital are included. The records are linked with other information about cancer registration and mortality.

The National Pupil Database (NPD) is held by the Department for Education and Skills. Every pupil attending a state school in England has a record on the database, recording the scores obtained on Standard Assessment Tests for each pupil. It also holds the name of the school attended and some characteristics about each school such as the proportion of children receiving free school meals. Schools are linked to local education authorities, and the NPD can therefore provide information at pupil, school and Local Education Authority level.

The Department for Work and Pensions (DWP) holds detailed information about the social security benefits and tax credits received by each individual. Historically, there have been a number of databases such as the Labour Market System (LMS), the Departmental Central Index (DCI), the Generalised Matching Service (GMS) and the Work and Pensions Longitudinal Study (WPLS). Increasingly, DWP data has been brought together to facilitate cross-benefit analysis. In addition, the Department holds data about individuals’ eligibility for and contact with various government services or programmes. This makes it possible to examine the effectiveness of government programmes designed, for example, to assist people to return to or retain work.

Her Majesty's Revenues and Customs (HMRC) holds information about each individual's receipt of tax credits and payment of tax and National Insurance.
Contributions. It also holds information about spells of employment. One such database is the National Insurance Record System (NIRS2).

One of the motivations for linking social surveys to government data is the increasing emphasis of the evaluation and monitoring of public policy and services. Indeed, administrative data themselves are increasingly being used to provide evidence for civil servants and public bodies and furthermore the design and development of administrative databases are being increasingly influenced by their potential value as an aid to research in their own right or in combination with linked survey data. For example, the recent creation of the National Pupil Database resulted from a decision to centralise records in order to allow the Department for Education and Skills to monitor the performance of individual schools and Local Education Authorities. This database has since been linked to a variety of surveys and used extensively as a research resource for longitudinal analysis in its own right. Similarly, the Department of Work and Pensions has undertaken a program of work to bring together separate administrative records to allow cross-benefit analysis – both current and longitudinal – and now provides the potential for linkages in a variety of contexts as well as for longitudinal analysis.

Given their nature and content, it is unsurprising that administrative data held by government departments and public sector organisations are among those most often linked to social surveys in the UK. However, administrative data held by private companies may also be exploited in this way. Some social surveys have made linkages to employer held information about earnings and occupational pensions. Market research surveys already link to commercial data, for example, about spending patterns and debt and there is great potential for using such data to address social research questions such as the relationship between consumption, diet and health, and between consumption and poverty.

3.2 Linkage to administrative data in the UK

The classic example of research that linked survey data and administrative records was provided by Richard Doll and A. Bradford Hill in their longitudinal survey of doctors’ smoking habits and their eventual cause of death (Doll and Hill 1954). The study established, beyond doubt, that smoking was a major cause of death from diseases such as cancer of the lung and coronary thrombosis. The impact of that one piece of evidence has still to be fully realised. As a demonstration of the value of linking survey findings with administrative data – in this case death certificates – Doll’s study has never been surpassed.

There were many factors in the study’s favour. The postal survey that was used to classify doctors by their smoking behaviour held a strictly limited number of questions and so was a negligible burden on respondents. The British Medical Association issued the first questionnaires on behalf of the study, and in later years it proved possible to trace a very high proportion of
doctors because of the accuracy and completeness of the Medical Register. As well as benefiting from a high initial response rate (69 per cent), the study experienced low attrition since the great majority of respondents appreciated its value. At that time there were no ethical committees and no obligation to seek consent for linkages, and the Registrar General provided administrative data on the cause of death of all medical doctors. Furthermore, the diagnoses of doctors’ illnesses are known to be more likely to be accurate than those for the general population. As a result, the study benefited from an accuracy and completeness of administrative data that would be hard to replicate.

In recent years, the linkage of individual level data from administrative records with data from survey respondents has become increasingly common in the UK, particularly on large-scale longitudinal surveys. Many of the major longitudinal surveys in the UK now encompass linkage to administrative data. A brief description of each of the major studies is given below. The linkages they carry out are summarised in Figure 2 below.

The Office for National Statistics’ Longitudinal Study (ONS-LS) and the Scottish Longitudinal Study (Scottish-LS) involves linking a sample of 800,000 and 274,000 people respectively from the Census of the Population and linking them to administrative health records.

The English Longitudinal Study of Ageing (ELSA) is a panel study of 12,000 people aged 50 and over and has linked to several administrative datasets in the fields of both health and economics field.

Although the British Household Panel Study (BHPS) does not currently link to administrative data, a methodological project based on a sub-sample of 1,000 BHPS respondents has recently linked to administrative economic data. This project is called ‘Improving survey measurement of income and employment’ (ISMIE).

All four of the British Birth Cohorts have linked to administrative data, primarily health records. Each of these studies follows a group of people born in a particular year. The National Survey of Health and Development (NSHD) comprises 5,000 people who were born in 1946. The National Child Development Study (NCDS) and the 1970 British Cohort Study (BCS70) follow people born in 1958 and 1970 respectively. The Millennium Cohort Study (MCS), the most recent of the cohorts, is tracking around 19,000 children born in the whole of the UK in 2000/2001.

The Families and Children Study (FACS) is a panel study of 8,000 families with dependent children and has linked to economic administrative records.

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study of 14,000 children born in the county of Avon in 1991/2 and has linked to health administrative data.
Whitehall II is a study of 10,000 civil servants which started in 1987 and which has recently linked to administrative health data.

As Figure 2 shows, linkage to administrative data in the health field is particularly common with economics and education linkages less widespread. This reflects the fact that central databases in the NHS have existed for longer than in other areas of social policy and many of the pioneering studies which have made use of administrative data have been in the field of epidemiology and public health, such as the example by Doll and Hill discussed above.

| Figure 2: The extent of linkages in current major longitudinal studies in the UK |
|---------------------------------|-------------------------------|
| Source                          | 1987                          |
| NHS                             | 1990                          |
| MDS                             | 1992                          |
| VLS                             | 1994                          |
| INLS                            | 1996                          |
| Health and Safety (HSE)         | 1998                          |
| PALS                            | 2000                          |
| ALSPAC                          | 2004                          |
| Whitehall II                    | 2006                          |

There are many other cross-sectional and longitudinal studies in the UK, not included here, that link to administrative data.

In some studies, the availability and nature of administrative data has been incorporated into the design of a longitudinal study and is integral to it (see Example Box 1 below).

Example 1: Estimating the impact of the New Deal for Lone Parents

The New Deal for Lone Parents (NDLP) is a national voluntary programme aimed at helping lone parents into or towards work and supporting them in employment. Personal Advisers offer work-related guidance through a series of contacts with participants. The quantitative evaluation of NDLP used administrative data in five ways (Lessof, Millar et al. 2003; Phillips, Pickering et al. 2003).

a) Two separate administrative databases (the Income Support Computer System and the NDLP evaluation database) were used to select a sample.
of approximately 70,000 eligible lone parents who had not participated in NDLP towards the end of 2000 and postal data was then collected from approximately 42,000 of them.
b) The NDLP evaluation database was used to identify approximately 1,800 of these individuals who participated in NDLP over the next eight months.
c) Data from the administrative databases and the postal survey was used to estimate a summary score which measured ‘propensity’ to participate in NDLP. These scores were used to match programme participants to non-participants and face to face interviews were then carried out with 1,250 individuals from each group.
d) Outcome data for each individual was provided by the face to face survey (entries into work, hours of work, wages etc.) and updated administrative data (which provided systematic information about exits from Income Support and receipt of other benefits). The rates of outcomes for participants and the matched non-participants were compared and the difference between the two provided the estimates of additionality.
e) Dolton and others continued to track the outcomes of the original sample of 70,000 using updated administrative data, refining the estimates of additionality through further analyses, and observing continuing outcomes to assess the longer-term success of the programme (Dolton, Smith et al. 2002).

4. Motivations for linking survey data with administrative data

The primary motivation for linking survey data with administrative data is to supplement or validate survey data by adding to or refining the substantive variables used by analysts to address their research questions. Another benefit of linking survey data to administrative data is to contribute information that is useful for improving survey quality in general.

Both of these motivations apply to both longitudinal and cross-sectional surveys but different considerations may apply when attempting to achieve these aims in the different types of study.

4.1 Enhancing Survey Data

Survey data can be enhanced by administrative data in several different ways. Most commonly, it is used to supplement survey data by providing additional substantive data items. Administrative data can be used as an alternative to asking questions directly to respondents. It may be that the information available from administrative data could have been gathered from respondents directly in the survey interview and the administrative data is used a direct substitute for survey questions. Collecting information from administrative records instead of respondents potentially frees up time in the interview for other questions or reduces the number of questions that need to be asked and hence the burden on the respondent. Often administrative data can provide information that would be difficult to obtain directly from
respondents in a survey interview. This may be because they are not able to provide the desired information or because there is likely to be recall bias in respondents' reports or because the recall task is too burdensome to be feasible. As discussed above, data on pupil attainment levels on Standard Assessment Tests (SATs) can be obtained from the National Pupil Database, data on hospital admissions from Hospital Episode Statistics or data on benefit receipt from DWP records. In examples like these, it may be more reliable to get this information from administrative data because respondents may not know or be able to report accurately SAT scores, exact dates and reasons for hospital admissions and exact dates and amounts of benefit receipts. Furthermore, some of the information that is available from administrative data e.g. the sub-division of state benefit payments into basic and supplementary elements or history of state pension contributions an extended period of many years may be extremely difficult if not impossible to collect reliably from respondents in a survey interview. Answers given by respondents in response to questions of this kind may be incomplete or inaccurate.

For these reasons, collecting data from administrative records may be a useful and effective way to supplement data collected in a survey interview and provide opportunities for research that may not have otherwise been possible. In addition it may be beneficial to respondents by reducing respondent burden, either by reducing the number of questions that need to be asked or by allowing some complex, detailed and uninteresting questions to be replaced by questions which respondents find more interesting or salient.

However, the major problem with supplementing survey data with data from administrative records is that it is not usually possible to obtain this administrative data for all survey respondents and this may mean that there is bias in any analyses using the administrative data if respondents for whom the data is not available differ from respondents for whom it is available.

There are three main reasons why it is not usually possible to obtain administrative data for all survey respondents. The first reason is related to the consent to linkage, the second reason is related to the success of the linkage and the third is related to completeness of the administrative data. These are discussed in turn below.

In the UK, written consent is usually required in order for the information held about a person on the administrative record to be released to a third party. Most surveys typically ask respondents to sign a data release form during the interview. Consent rates for the use of administrative data are usually less than 100 per cent of survey respondents. Non-consenters may therefore introduce a potential source of non-response bias. Evidence from the UK Millennium Cohort Study in relation to birth registration data shows that although consent rates were high (92%), some groups were less likely to consent than others. In particular, consent was lower among mothers living in
Northern Ireland and mothers from ethnic minority groups (Tate, Calderwood et al. 2005). A methodological study based on the British Household Panel Survey (BHPS) also revealed evidence of consent bias showing that respondents who were aged 40-49, or single householders, or had shorter previous interviews or had problems with that interview were less likely to consent (Jenkins, Cappellari et al. December 2004). The ethical and legal issues in relation to consent in the context of longitudinal studies are discussed further in Section 5.

Furthermore, even respondents who give consent may not be successfully matched to the administrative data, and again it is possible that there will be systematic biases in the characteristics of non-matched cases (for example if matches cannot be found for women who have reverted to their maiden names following a divorce). The process of linking survey data with administrative records usually involves a form of probabilistic matching using identifying variables which are available in both the survey and the administrative datasets. ISMIE, the methodological project based on a sub-sample of BHPS respondents, reported overall match rates of around 74% of those giving consent and that there were some differences between matched cases and those who consented but were not matched (Jenkins, Lynn et al. October 2004). The technical issues in relation to probabilistic matching are discussed further in Section 5.

Finally, the administrative data may not be available for all survey respondents i.e. there may be under-coverage in the administrative records. The most obvious UK example is that databases such as the National Pupil Database and Hospital Episode Statistics only cover state provided services and does not include the minority of people who receive private education or private health care in the UK. Hence, surveys relying on obtaining estimates of pupil attainment or hospital care from administrative records must either accept that this information will not be available for this non-random group of respondents, and hence there will be bias in the survey estimates or attempt collect this information in another way. Most commonly this is done by collecting data directly from this sub-sample of respondents which is only a partial solution to the problem of bias as it means that there is non-random variability in the source and reliability of the data.

However, the problem of non-response bias affects all estimates based on survey data. All surveys suffer from unit non-response and particular data items suffer from item non-response. As a result, it may be thought that the problems of potential bias in linked administrative data may be trivial compared with overall non-response. Moreover, a great deal of information is typically available from the survey data about respondents for whom administrative data cannot be obtained and this information may be used to correct for potential bias in the administrative data through for example, weighting. Moreover, because survey questions are usually an extremely scarce and fairly expensive resource, collecting data through linkage may still
be an attractive and cost-effective prospect despite this problem of potential bias.

If concerns about potential bias are not considered serious, then a potential problem which is more substantive in nature should be considered. This is that the substantive variables that are available from administrative data sources may not always be reliable i.e. there is measurement error in the administrative data. For example successive evaluations have shown that a minority of individuals classified as participants in government programmes have no recollection of taking part. Some of these anomalies may be a consequence of a bureaucratic definition of ‘participation’ which is not meaningful to the individual, some may reflect over-enthusiasm of officials who may have incentives to enrol programme participants, and some may be simple administrative or recall error. Inevitably, some data entry errors will be made by officials, particularly where the accuracy of a particular field or variable does not visibly effect the action that is being carried out, such as a benefit payment or hospital admission. If there is measurement error in the administrative record, then it may not be advisable to rely on record linkage to substantive data items. Perhaps respondents are more reliable than records after all?

Of course, unfortunately neither respondents nor administrative records are reliable sources of perfect information. For this reason, in practice administrative data is often used not to supplement or substitute for information collected directly from respondents but as a way of validating survey data. Particular substantive variables are collected both by putting questions to respondents and from administrative data and the results are compared. For example in the UK Millennium Cohort Study, data on birth weight was obtained from birth registration data as well as from the mothers who responded to the survey, and the different estimates obtained were compared with each other. This analysis revealed a high level of consistency between the different sources. The birth weight reported in the survey was within 100g of the registration weight in 92% of cases (Dezateux, Foster et al. 2005). In this example, the linkage to administrative data has meant that analysts are able to have more confidence in the reliability of the survey estimates and hence in any substantive conclusions based on these data. Furthermore, where such validation suggests that the administrative data is reliable, it can be used to minimise the problem of item non-response in the survey data. Where the missing item is available from administrative data it may be possible and appropriate to impute information that is missing in the survey data directly from the administrative record. In this example, if birth weight is refused or not known by particular respondents, it may be possible to substitute the information on birth weight available for that person in the birth registration data.

Administrative data can be useful in dealing with item non-response in the survey data, even if the missing data item is not available from the administrative record. Where administrative data is reliable and complete,
variables from administrative records may be used in addition to or instead of variables from the survey data as auxiliary variables in classic imputation methods such as conditional hot-decking.

In other examples, the validation of survey data using administrative data may show much lower levels of consistency between the different sources of data. This may mean that is appropriate for analysts to show more caution in their research results and in the future may lead to improvements of measurement techniques, in both the survey and the administrative records.

Finally, administrative data can also provide measures of substantive variables at additional time points to the survey data collection. It is possible to enhance survey data by linking to historical data from before the start of the survey (such as employment or pension histories), variables relating to episodes in between the times of data collection (such as detailed benefit histories) or data relating to spells or events after the survey has finished or after the respondent has left or been lost to the study (such as exam results or date and cause of death).

To conclude this section, there are particular issues regarding enhancing survey data which are particular to longitudinal surveys. In general, as longitudinal studies follow the same individuals over time, they will generally wish to carry out repeated linkages different points in time to continue to update their linkage data in the same way that they update their survey data.

Measurement error in administrative data can cause more problems in a longitudinal survey than in a cross-sectional survey, where the main intention is to provide prevalence estimates for populations or sub-groups or to explore relationships between variables across the population. While random measurement error at an individual level may not detract from these population estimates, for longitudinal surveys the main intention is to look at change over time at an individual level. This means that measurement error at an individual level can often have a greater impact on the analysis of change over time.

\[4.2 \text{ Enhancing Survey Quality}\]

Although the primary motivation for linking to administrative data is usually to enhance to survey data, there are also many potential benefits for the survey in general that can flow from such data linkage. The administrative data may provide substantive information that can be used to help better understand survey non-response and may be used for assessing the validity of key survey estimates. In addition, the administrative data may be able to provide contact information for survey respondents which may help improve survey response and representativeness. Finally, the inclusion of administrative data linkage in the survey may also influence respondents’ perceptions of the study in a positive way. These are discussed in turn below.
Administrative data can also provide information to help improve the understanding of any potential biases in survey non-response and, if appropriate, to deal with this by developing weighting strategies (see Example 3 below).

In the context of a longitudinal survey, it is particularly important to understand potential bias that can occur as a result of attrition between waves or sweeps of data collection.

**Example 3: MCS as an example of understanding survey non-response bias**

The sampling frame for the Millennium Cohort Study was drawn from Child Benefit records of the Department of Work and Pensions, and this administrative data could then be used to analyse losses from the original sample of families at the initial sweep of data collection. A proportion of the sampled families (10%) had been excluded before fieldwork commenced, mainly as a result of the exclusion of ‘sensitive cases’ by the Department of Work and Pensions and opt-outs by the families themselves. As the administrative data was available (at aggregate level) for all families in the original sample it was possible to compare the characteristics of those families who were excluded at this stage with those families who were not.

It was found that claimants paid by order book, older claimants (34+) and large families (4+ children) were more likely to have been excluded from the sample than claimants paid by bank account, younger claimants and smaller families. For families that were included in the survey, this information was available at individual level so it was possible to look at non-response to the survey in a similar way. As well as some country and ward level differences in response, it was found that claimants with a title ‘Miss’ were less likely to respond, whereas claimants who were over 33 and paid through a bank account were more likely to respond. This information was used to develop non-response weights that could be used by analysts. In future sweeps it may be possible, in similar fashion, to use child benefit records to devise weighting for families who withdraw or are lost to the study.

(Plewis 2004)

Administrative datasets that are population references can be used to inform estimates of under-coverage of the survey and information on attrition by comparing the characteristics of the survey sample with population estimates obtained from the administrative data. This is possible even if the administrative data are not used as sampling frames.

By improving the understanding of non-response bias and attrition, the linking of longitudinal surveys to administrative data can thus lead to substantial improvements in survey quality.

An administrative dataset may also be used to produce population estimates of key variables which can then be compared with the estimates obtained by the survey. For example, for a survey of state school pupils in England,
estimates of attainment obtained from the survey could be compared with population estimates obtained from the National Pupil Database.

Another benefit is that administrative data can be used to help keep in touch with respondents. For example, the administrative data may contain home addresses of respondents and other contact information. Again this is particularly important in the context of a longitudinal survey, as mobility can be a major contributor to attrition. In the Millennium Cohort Study address updates are provided before each sweep by the DWP, in order to help minimise the loss of sample through mobility. Other administrative datasets that have been used to provide addresses for respondents are the National Health Service Central Register which holds the name of the health authority in which every person registered with a GP in England is resident and the Driver and Vehicle Licensing Authority which holds the home address of everyone licensed to drive in the UK and the registered address and owner of every car in the UK.

Another potential advantage of linking longitudinal studies to administrative datasets may flow from the way in which this process influences respondents’ perceptions of the survey. Asking permission to link to administrative databases that are perceived as ‘important’ or ‘official’ may enhance the status of the survey in the eyes of respondents and hence increase their loyalty or commitment to the survey. In the context of a longitudinal survey, this may make them more likely to co-operate and less likely to withdraw from the study. However, the reverse is also true; data linkages may have a negative effect on their perception of the survey and their future co-operation, especially if they feel that collecting data from these administrative sources is intrusive or inappropriate. At present, there is little available UK evidence about whether including such linkages has a positive or negative impact on the survey as a whole.

5. Practical Issues

Although administrative data linkage may be perceived as a relatively economical and straightforward way to enhance surveys, the actual process of linking surveys to administrative data can be costly, often time-consuming and complex. In general, most linkages involve obtaining legally-binding informed consent from survey respondents, providing the holders of the administrative records with individual level data which they can use to identify the appropriate records in the administrative data, processing and documenting the administrative data and managing its release to users. This can involve considerable additional resources and should therefore be embarked upon only if these issues are understood and budgeted for.

There are also decisions and priorities which should be agreed upon before any study is started. These include the question of who should be given access to administrative data and how can confidentiality be assured - particularly as the providers of data may still hold the original records that
would enable matching back to personal details. There may also be concerns about the degree of regulation that would be required and whether this would add to recruitment difficulties or inhibit long term studies. Some linkages may need personal details such as name and date of birth, to link across different data sets. Will these be available and if not, can a reliance on probabilistic matching meet the criteria for a viable study.

**Ethical and Legal Issues**

There are ethical and legal considerations to be resolved in linking longitudinal surveys to administrative data. Most linkages require the written consent of respondents for the administrative data to be released. This usually asked respondents to sign a consent form. Before seeking to gain consent from respondents, researchers must first agree with the holders of the administrative data and their legal advisors a form of words to be used on the consent form. For surveys in which consent is being given by parents for the release of data about their children, it is also necessary to ensure that consent is collected from a parent who is legally able to do so, for example, consent from a step-parent or even a non-married natural father may not be legally valid. As well as the legally required minimum, ethical considerations mean that it is common to include additional information on the consent form to ensure that the consent is informed. For example, explaining to respondents what the administrative dataset is, the nature of the information that will be obtained and the reasons for requiring this linkage. An appropriate balance must be struck between the desire to maximise consent rates in order to minimise non-response bias and the ethical requirement – present in all surveys - to ensure that consent is fully informed. This applies particularly to longitudinal surveys, in which the trust which is built up during a long term relationship may mean that respondents often try to be compliant even if they are unsure about the exact nature of the data linkage. This increases the obligation to ensure that the respondents are fully informed about the consent that they give.

Ethical and legal considerations have been taken further than this in the approach taken by the English Longitudinal Study of Ageing (ELSA). Respondents have so far been reminded at each interview of the permissions that they have given and the implications of these permissions. Following the same approach, it could be argued that when respondents withdraw from a longitudinal survey they should be reminded of the permissions they have previously given, especially if this allows the survey to collect information about them into the future, in order to give them the opportunity to withdraw these permissions as well.

The provision of additional data from administrative records could also increase the risk of disclosure, especially in longitudinal surveys where the cumulative nature of the data about individuals means that disclosure risk is already higher than in a cross-sectional survey. This is one of the reasons that administrative data is so safely guarded. In some cases, administrative data may contain information that respondents do not know themselves and may
not wish or need to know. It might be suggested that for this reason too, administrative data should not be publicly available or that access should be restricted, for example through the use of special downloads or safe settings.

**Technical**
The actual mechanism for linking survey and administrative data involves identifying and extracting each individual's record or records from the administrative database. Identifying survey respondents in administrative records can sometimes be done using unique identifiers such as National Health Service or National Insurance number but more commonly involves probabilistic matching using individual level characteristics such as sex, date of birth and home postcode.

For any probabilistic matching, there are issues to consider in relation to what the matching variables should be and what the criteria for a valid or acceptable match should be. The aim of any matching process is to minimize the errors which occur when administrative data are matched to survey data for the ‘wrong’ case (‘false positive’) or when there should have been a match but none is made (‘false negative’). The methodological project based on the BHPS sub-sample reports overall match rates of around 74% of those giving consent and estimates false positive rates of around 0%-10% and false negative rates of around 5% (Jenkins, Lynn et al. October 2004).

The consequences of such false positive matches may be more problematic in a longitudinal survey that tracks the same individuals over time than in a cross-sectional survey. Similar problems may arise where the matching process uses apparently unique identifiers such as NHS number or NINO which can be reported or recorded in error, or can be duplicated.

### 6. Conclusion

Linking longitudinal surveys to administrative data is an important development in longitudinal research and can enhance both the substantive data and survey quality. By validating and supplementing survey data, administrative information can improve data quality and provide additional research opportunities for analysts.

The procedures involved in linkage do, however, present new challenges to data holders and analysts over the methods to be used in the matching process for obtaining the administrative data and minimizing potential bias and measurement error, while also finding ways of managing the disclosure risk.

The potential benefits to the survey quality itself are often specific to longitudinal surveys. The ability to use administrative records to track respondents and to understand and adjust for attrition to longitudinal samples are important benefits which can ultimately improve the quality of longitudinal surveys and longitudinal data.
Finally, and especially in longitudinal surveys, there are important legal and ethical issues to be considered when this type of linkage is used.

References