Administrative vs survey data for longitudinal analyses

Paul A. Smith
Ann Berrington
Peter W.F. Smith

S3RI and Dept of Social Statistics & Demography, University of Southampton

July 2019

This work was funded by the Economic & Social Research Council under grant no. ES/T001038/1.

Administrative vs survey data for longitudinal analyses

Paul A. Smith, Ann Berrington, Peter W.F. Smith
S3RI/Dept of Social Statistics & Demography, University of Southampton

Executive summary

Social longitudinal analyses in the UK have largely been based on survey data, assisted by the investment on cohort and panel studies. In recent years there has been a move to utilise alternative data sources and particularly administrative data. The differences in properties in and benefits and drawbacks of the different types of data are summarised.

Administrative data sources are becoming more widely used in a range of contexts, but it is more challenging to discover the detailed changes and metadata. It is recommended that the possibility of commissioning 'administrative data biographies' should be explored, which will describe the evolution of the administrative sources and the way the information in them is gathered.

Linked administrative data and combined administrative and survey data present a wider range of analytical possibilities, but there are challenges in consent for linkage and in undertaking the linkage itself.

A spine is a list of people derived from one or more administrative sources, and to which other data can be linked. We describe some of the considerations in constructing and using a spine in support of longitudinal survey taking.

1. Introduction

In recent years there has been a strong development of approaches based on alternative data sources – administrative and transaction data, and in some cases 'big data' ("found data" in the sense of Connelly *et al*. (2016), because it is not designed for statistical purposes) – in official statistics and major research. For convenience, we will refer to these sources generically as *administrative data* here. The main drivers for the use of administrative data have been financial, since they are potentially available at much lower cost than survey data, but have also been driven by their wider coverage, big increases in data availability within government, and increases in computing power to process large datasets. Development of new methods, and experience with implementing them in practice, have been going on in tandem. In the UK this has been supported by the development of research infrastructure such as the Administrative Data Research Network, but also been mirrored in the development of statistics in government to make greater use of new data sources, for example in the ONS's Data Science Campus.

While there is considerable literature comparing the merits of longitudinal data with cross-sectional data, and of administrative data with survey data, there is little comparing longitudinal survey data with longitudinal data from other sources. However, Hurren *et al*. (2017) discuss the advantages and challenges of using longitudinal administrative data with reference to child maltreatment.

In this review we aim to summarise recent thinking on the potential, advantages and challenges of using survey data and administrative data (separately and together) for longitudinal research.

2. Survey data

Survey data have been the default data for many types of longitudinal analyses, aided by the existence of some long-running longitudinal surveys and their associated datasets. They present a number of advantages, but also some attendant disadvantages, summarised below.

2.1 Data definition and collection

While concepts and measurements are under the control of the survey takers, there are many challenges in collecting survey data, including: defining the concept to be measured; developing and testing questions which capture the information required to derive that concept; and controlling the interview process to maximise the response and minimise any measurement error. A longitudinal survey may still need to resolve conflicting user requirements and cost-benefit trade-offs, but once the requirements are defined they can be implemented. This makes the data more *relevant* for answering specific questions of research and policy interest, and is expected to keep some non-sampling errors (such as measurement error) low, improving *accuracy*.

In a single wave of a survey all the variables are collected at the same time, which means that associations are less likely to be contaminated by timing differences. It also allows simultaneous collection of biomarker data (Ruiz *et al*. 2017).

There are still some challenges – changing social contexts or developments in research agendas may alter the requirements, and hence the availability and comparability of data. The measurement occasions are tied to a defined pattern of survey waves, so they measure the state of each variable at the time of the survey, with transitions between states either inferred by comparison across waves, or identified by recall with its attendant risks of bias through telescoping (Tourangeau 2018) or seam bias (Callegaro 2008).

The repeated collection of data from the same respondents potentially gives very rich datasets, but the respondents may be influenced by the survey process, which may affect their responses to the survey in different waves (rotation group bias, van den Brakel et al. (2020, forthcoming)), or change their behaviour, known as panel conditioning (Warren & Halpern-Manners 2012).

Data may be collected in a number of modes. Face-to-face interview mode is generally accepted to be the most accurate, but it is expensive because of the cost of the interviewer. Internet and post are much cheaper, but suffer from nonresponse and (for internet) break-off.

2.2 Defining a population and sampling

In a longitudinal setting it is always challenging to define the population about which inferences can be made, since the survey covers many periods (Smith *et al*. 2009), but this is more acute in surveys where the initial sample is selected from a frame from a specified time. However, the choice of frame is again under the control of the survey designer, although the frame is often itself an administrative source, which blurs some of the distinction between surveys and administrative data. For some population definitions, methods are needed to update the sample to include people who were not present when it was initially selected (such as births or immigrants). A boost may be added (Watson & Lynn 2020, forthcoming), but the additional sample often complicates data management and analysis. The following rules for the survey may allow some elements of the population to be updated (such as births to original sample members to represent births in the population), but will not cover all cases. Longitudinal datasets derived from rotating panels, eg the Labour Force Survey, which drop respondents after a fixed time and therefore allow the sample to change with the population, are less prone to this, but do not provide consistent long-term data on the same individuals.

Sample design for a longitudinal survey is a trade-off among competing requirements for accuracy in different variables and subpopulations, costs, interview length and predictions about recruitment and attrition rates. This is not the place to attempt a review of design considerations (see Smith *et al*. (2009) for an overview), but note that differential sampling rates to meet various requirements can complicate the analysis of the collected data, through use of weights, or through combinations of different types of samples (eg on Life Study, Dezateux *et al*. 2016 which had a nationally representative probability sample and location-based samples). There has been some debate about the necessity for weights in longitudinal analysis, see WP5 and Goldstein *et al*. 2015).

2.3 Maintaining the data

There are costs in maintaining a longitudinal survey sample. To support repeated data collection, sample members must be tracked, typically by tracing activities and keep in touch (KIT) communications between waves. Even with these efforts there is still attrition, and in severe cases high attrition could lead to longitudinal studies being stopped (eg the National Longitudinal Survey of Children and Youth (NLSCY) – a Canadian study – was halted when it became apparent that it was not representative). Data have to be linked across waves in a suitable way, and item nonresponse must be dealt with by a suitable longitudinal imputation approach.

Editing, to identify and correct errors, is also an important activity, and is easier in a survey where the data collection is under the study's control, so that unexpected observations can be checked at the time of collection. A survey, even with a large sample, will also have many fewer records to examine than an administrative dataset. In extreme cases the cost of cleaning a complete administrative dataset may be so great that it is only practical to clean a sample, and use the cleaned sample in further processing.

2.4 Ethics and confidentiality

A survey approach deals directly with the ethics of data collection and analysis, because respondents are asked for their consent to participate in the survey (Lessof 2007). Analytical outputs are also almost automatically protected against disclosure of any respondent's attributes by the sampling, which makes it less likely that any records which are unique in the sample will also be unique in the population. Nevertheless the large number of variables available in longitudinal microdata means that they continue to be very sensitive (that is, susceptible to being identified), so microdata and low level geographic data have restricted access.

2.5 Summary of benefits of longitudinal analyses based on survey data

The key benefits of surveys as a source for longitudinal analysis are the ability to collect the desired target variables (including biomarkers if necessary) from a defined population with high quality. The quality of the data should feed through to the quality of the analysis and inference, although there are technical challenges to be overcome in the process. The ethical basis and confidentiality procedures are straightforward.

3. Administrative data

Administrative data offer the prospect of large amounts of information at low marginal cost, and there has been a sustained development of these resources for statistical purposes (e.g. HM Treasury (2014)). Their principal advantage is that they have no marginal collection cost (though there may be other costs), but there are other properties which interact with longitudinal analysis as described below. In particular, administrative data systems are susceptible to changes driven by changes in administration, policies and systems, and it is important to filter out these changes in any analysis. In order to understand an administrative dataset and effects of changes in policy and context on the data contained within it, Connelly *et al*. (2016) call for a "biographical understanding of the administrative system". It seems that many researchers in different fields are trying to understand the properties of the same administrative datasets, but that there is no systematic documentation and evaluation of these sources at the most detailed level. One area where ESRC could improve the UK data infrastructure would be to fund construction of 'administrative data biographies', descriptions of the properties and metadata of the systems and when they have changed, for the major sources. This would necessarily involve the cooperation of the bodies which produce them, but could include, for example, details of the instruments used to collect administrative data and when they change, as well as changes in policies and practices.

Recommendation: Explore the possibility of funding construction of 'administrative data biographies'.

3.1 Data definitions

Administrative data requirements are defined to support administration, so the variables included are determined by that requirement. Where these are the variables of substantive interest, they will be measured very well by the administrative data (much better than in self-reports); for example lifetime earnings would be much better measured by a tax system than by recall, and medical histories by GP and hospital records. Administrative data are also unobtrusive, in that they record events close to the time they happen, and do not require a respondent to recall potentially uncomfortable or traumatic situations.

It may be possible to influence the controllers of administrative data to include one or two key variables, particularly identifiers which enable datasets to be linked (see section 3.4) and therefore enable a wider use of the administrative data. Additional collection has a substantial additional cost in the administrative process because it is implemented for the whole population, and so would need a strong cost-benefit justification. Hence, in general only the administrative variables are available. These are not controlled by the researchers, and are susceptible to changes in the underlying administrative systems, which may change or remove variables unexpectedly (recent examples include the introduction of Universal Credit, changes to rules for Free School Meals, and changes to Child Benefit entitlement; see also van Delden *et al*. (2011)).

One important variable which is often not available in administrative data is an identifier showing which people constitute a family or household. Households are often important in analyses, and it is challenging (and therefore subject to error) to reconstruct these purely from administrative variables (Zhang 2011); errors may be associated with characteristics which define groups of particular policy interest, such as those with complex living arrangements, and may therefore disproportionately affect some types of analysis. Identifying families is important from a 'linked lives' perspective which says that the development of an individual is intertwined with development of other family members.

Administrative data are able to record changes at any time, and thus present more granular event data, not susceptible to recall or similar biases. This should improve event history analyses (possibly with further adjustment to deal with censoring, see Courgeau & Najim (1996)). The period over which people interact with administrative data may however be fuzzy, with lags in identifying changes possibly long; Zhang & Fosen (2012, Table 2) document an example where some changes are still taking place 7 years after the reference date. Administrative data can be collected with some delay relative to the reference period (eg income tax), but timeliness is not usually a critical factor in longitudinal studies. Long lags though may have an important effect on estimates of changes and gross flows.

3.2 Coverage

Administrative sources cover the population of people who interact with an administrative system. If this population coincides with or approximates to the one required for analysis, then the administrative sources provide data for the whole population. This particularly means that population subgroups will be well covered, so administrative data increase the scope for analyses of these subgroups.

Administrative data can however suffer from under- and over-coverage of a population of research or policy interest. The use of administrative data as the basis for sampling means that surveys may also suffer from undercoverage; in either case a separate exercise is needed to measure undercoverage, basically a problem of population size estimation (many details of which are covered in the chapters of Böhning *et al*. 2017). Adjustment for undercoverage is usually achieved through weighting with weights calibrated to 'known' population totals which have been adjusted for undercoverage. Overcoverage principally affects administrative data, and means that records that should not be included are present in the data and affect the analyses. Zhang (2015) offers some strategies for modelling overcoverage, but this always requires a source without overcoverage.

Administrative data are not affected by attrition. In order to continue to gain the benefits of using an administrative process (or avoid the sanctions from failing to use it), people need to continue providing data in support of the system's administration. However, people may move in and out of the population covered by the administrative process, and there may therefore be incomplete histories.

In principle the administrative data cover all, or a large part, of the population of interest, and therefore facilitate the analysis of subpopulations which may be included too infrequently in samples to allow separate analysis.

3.3 Ethics and confidentiality

The ethical basis for use of administrative data is more challenging than for survey data. Sexton *et al*. (2017) highlight the competing public goods of data re-use to improve efficiency and outcomes, and privacy. They also show that where there is a lack of trust in the system, boycotts of data collections (eg the schools census) can result, which would affect longitudinal analyses.

Because they contain all records, administrative data also present a higher risk of disclosure, and are subject to more stringent security and confidentiality protection. Gaining access may be challenging in some cases, and the data are usually available only in secure environments. Analyses may need to be assessed for disclosure risk before they can be released from these environments and published.

3.4 Linkage and linked administrative data

With an administrative dataset there is a need for linkage across time to produce a longitudinal dataset for analysis, and also possibly across sources to expand the range of variables. Within an administrative system there will generally be a consistent identifier which will provide high quality links. But linkage compounds any coverage problems, which now have components deriving from two or more periods and/or sources. Any coverage errors potentially affect the linkage, and any linkage errors in turn may affect longitudinal analyses. Linkage of two different datasets also raises additional ethical questions, about whether participants have an opportunity to consent to linkage (and if they do how much the withholding of consent produces a selection effect on who is included in the linked data; see also 4.1), and the greater number of variables in the linked data, which make them more sensitive.

Administrative datasets potentially contain variables that have been measured at different times, and this may affect the estimation of relationships. This property is expected to be more prevalent in data linked from multiple administrative sources, where there is generally no particular reason for measurements to be simultaneous. We did not find any empirical evidence of the extent of timing differences in variables in administrative data or their impact on analyses, and this is an area where further investigation would be worthwhile.

Recommendation: Commission some example studies which investigate the effect of timing differences in single and multiple-source longitudinal administrative data on analyses.

3.5 Analysing longitudinal administrative data independently

Some variables will be relatively well captured by administrative data, which may therefore be an effective source for longitudinal analysis, with the benefit that most of the population is covered. The the concept in the administrative data may not be exactly what is required, and there may be lags in obtaining, cleaning and processing the information relative to a survey. The best process for gathering administrative data is an open question. In principle they can be available continuously, but this requires updating and potentially large storage, unless data are stored as changes between states with an associated timestamp. It is likely to be better to have periodic snapshots, and to ensure that appropriate editing/cleaning processes are applied. The UK Statistics Authority has good guidance on cleaning and assessing the quality of administrative data (https://www.statisticsauthority.gov.uk/osr/what-we-do/systemic-reviews/administrative-data-and-official-statistics/). Although cleaning is

important for analytical variables, Randall *et al*. (2013) note that it may not be useful for linkage, because while it improves linkage rates for (true) matches, it also increases linkage rates for non-matches.

The coherence of the administrative data with the concepts being analysed is important – how far do the administrative data provide the right variables to answer questions of substantive interest? And this is particularly important if the administrative data are being considered to replace survey data rather than supplement it. There will always be pressure to make such replacements, for reasons of cost and coverage. Van Delden et al. (2016) distinguish four cases – control, accept, adjust, reject, depending on whether there are no, minor, adjustable or nonadjustable differences respectively in concepts.

The Longitudinal Studies in Scotland, N Ireland and England & Wales already combine information from successive population censuses and administrative data sources for a sample of their respective populations, and therefore provide an important source of longitudinal information, though on a more restricted range of variables than the main ESRC-funded longitudinal surveys. Nonetheless, these are suitable for many research questions and are widely analysed (see https://calls.ac.uk/outputs/journal-articles/).

3.6 Summary of benefits of longitudinal analyses based on administrative data

The key benefits of administrative data as a source for longitudinal analysis are the low (marginal) cost, and the high coverage, which allows much more detailed breakdown and analysis of population subgroups. Where the variables of interest are the administrative data variables, the quality of the data will be high, but the range of variables is generally smaller than for surveys. Administrative data are typically collected continuously, so analyses requiring good information on the timing of events may be of higher quality.

4. Combined administrative and survey data

The advantages of survey and administrative data complement each other well, so combined datasets have become popular (see, for example, Meyer & Mittag, 2019). They do indeed offer benefits, but these are offset by an additional challenge which is the need to link data sources, and the dependence of analyses on the quality of linking and the linkage error. Since the linkage uses a survey as one source, the linked data consist of only a sample of units and are therefore subject to the analytical challenges derived from surveys, such as weighting and nonresponse adjustment. A specific application in the use of an administrative spine to support longitudinal surveys is discussed in more detail in section 5 below. See also WP3.

4.1 Linkage and consent

Combinations of administrative and survey data for longitudinal analysis usually require linkage of the microdata, so that associations between variables can be modelled. There are several stages in the process of moving from a frame to a sample linked to an administrative dataset, and all are subject to some type of error (see Sakshaug & Antoni 2017 for an overview).

There are many procedures for linkage. An ideal solution is to collect a suitable identifier during the survey. However, in some cases sources must be linked without a common identifier, and then some comparison of the data is needed, coupled with probabilistic matching. Where high quality of matches is important (as it is in longitudinal surveys where all the cases have a high value so that errors should be kept to a minimum), clerical (human) review of challenging cases will be needed.

When a survey is involved, respondents' consent should be sought to link their data to administrative records, and this is not always forthcoming. This can have a large impact on the data available for analysis, since consent may be given differentially in different groups, presenting a situation not unlike (and additional to) non-response bias. This means that analyses of variables related to the variables

on which consent is differential will be biased, unless there is a suitable source for adjustment; often there isn't.

Sakshaug & Kreuter (2012) find that consent rates vary substantially depending on the survey context (between 24% and 89% across a range of surveys in different countries). The impact of differential consent in determining which characteristics are over-or underrepresented in the final analysis dataset also varies.

Sakshaug & Huber (2016) consider consent in a panel study, and find that "linkage consent bias […] decreases over time when respondents who do not provide consent in a prior wave are asked to reconsider their decision in subsequent waves". Sala *et al*. (2014) similarly discovered that there is more chance to obtain consent from previous non-consenters if their previous decision is not mentioned in the interview, which presents a ratchet-like approach to increasing consent over the waves of a longitudinal study. See also Mostafa (2016) for a discussion of consent bias.

4.2 Imputation

One potential use of linkage is to derive good imputed values in the case of nonresponse to a survey. This would by most useful where a survey is the only vehicle for collecting good quality information about the concept of interest, but where variables in the administrative data are good predictors for this variable. After a refusal to participate in a survey it would not be ethical to obtain a good prediction of essentially the same information that had just been refused, by linkage to an administrative source, but in cases of noncontact it is more reasonable to use the administrative data as the basis of an imputation.

4.3 Summary of the benefits of longitudinal analyses based on combined survey and administrative data

The key benefits of linked administrative and survey data as a source for longitudinal analysis are the availability of variables from high quality data collections under the control of the survey takers coupled with high quality administrative variables, and the ability to use correlated administrative variables to understand the coverage and biases in the survey data.

5. An administrative data spine

There is a substantial range of administrative data sources (e.g. ONS 2016), much derived from government activities and held by different government departments, but also a range of private sector sources. The ONS has been investigating the use of administrative data sources in the construction of a population spine, as the basis for development of population estimates without the need for a population census, and this has led to repeated deliveries of experimental outputs (https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject), and New Zealand has also been undertaking similar research (Black 2016). A spine is not a population register, as there is no requirement for people to register and update their entries in a timely fashion. It is a list constructed from one or more sources to which other sources can be linked.

The ESRC longitudinal studies review (Davis-Kean *et al*. 2017) had as one of its major themes the use of administrative data to support longitudinal survey taking. This is closely related to the research infrastructure development in the Administrative Data Research Partnership (ADRP), which aims to maximise the potential of administrative data as a resource for high-quality research in the UK by making datasets available.

Administrative data have several advantages; they generally have high population coverages (depending on the type of data), and this means that there are larger numbers of records for minority groups. They also present a number of challenges, summarised in Hand (2017).

Access to data sources containing identifiable records for the construction of a spine is, however challenging. Procedures for data sharing within government have become progressively easier with the Statistics and Registration Service Act 2007 and Digital Economy Act 2017, but there is a continuing view that the risks from a disclosure of personal data could have very serious consequences for data collection in the UK statistical system. Data owners have therefore taken a cautious approach, and it is challenging to assemble the variety of data to form a spine, with appropriate safeguards for confidentiality and secure storage and access.

The vision at the outset of the Administrative Data Research Network (ADRN) was that a wide range of administrative data would be available and could be linked to generate bespoke linked datasets for analysis of particular variables available from at least one of the sources. This process did not work well in England, where there has not been enough agreement to exchange of data because of the risk of a reaction from the public. In Wales and Scotland where the datasets involved are smaller and the devolved administration means that most of the data originate from devolved institutions under more connected leadership, it has been possible to produce a spine, and to use this as a basis for producing linked analysis datasets. See, for example, the Welsh Secure Anonymised Information Linkage (SAIL) annual reports available from https://www.healthandcareresearch.gov.wales/secure-anonymised-information-linkage-databank/ and the Scottish electronic Data Research and Innovation Service (eDRIS) website: https://www.isdscotland.org/Products%2Dand%2DServices/EDRIS/.

Constructing a spine is a challenging procedure even when the data are all available. The UK does not have a unique personal identifier, and although there are various identification numbers (of which the National Insurance Number (NINo) and NHS number are the most pervasive), they are not usually available on datasets to form the basis of linkage. Therefore, much linkage needs to take place with name, date of birth and address, and these variables are in some cases subject to incompleteness and measurement error which affect the ability to link them. In these cases some form of probabilistic linkage is needed for at least some of the records, with associated potential for false or missed links.

The ONS approach has been to use four administrative sources, two of which (HESA and Pupil Census) cover specific age groups, and to consider the linked records to represent part of the population of interest only when they derive from more than one source. This leaves a substantial group of records deriving from only one source whose status is not clearly known. We can interpret these using a "signs of life" approach, as described in the Irish context by Dunne (2015) and the Estonian context by Maasing *et al.* (2017).

5.1 Spine coverage

The constitution of a spine is not straightforward, and there are several considerations about what should be included:

a. unit coverage: a population spine used as a framework for linked data should include all the people who appear in any of the datasets to be linked. This is a maximal view of a spine. All the people who interact with an administrative system will be included, with sufficient identifying information to allow other datasets (administrative or otherwise) to be linked to the spine. Such a spine will suffer from overcoverage from duplication, missed links and overcoverage (eg of emigrants who have not de-registered or been subsequently removed by list cleaning exercises) in the base datasets. If the spine is intended as a basis for defining a population, for example to investigate the representation of a particular subpopulation in a longitudinal survey recruitment, then a process for removing overcoverage will be needed so as not to produce biased results from differential overcoverage of this subgroup.
b. temporal coverage: a spine can be dynamic, with the latest information always used to update it (and this is the principle of business registers (College 1995, section 2.5)). However, in a longitudinal framework it might be better to be able to track changes through time, which would suggest a longitudinal spine, containing different information for the same units at different times. Such a resource would however be much more complex to construct and maintain; it would certainly require records to be linked over time, adding another process

which is potentially subject to linkage error. Contradictory information from two sources at approximately the same time might indicate a change (e.g. of address), but not give any clear information on when it took place, for example. The more information is included, the more sensitive the spine will be and therefore the more stringent the procedures for accessing and using it.

c. identification information: identifiers or matching variables are needed for linkage of units in different source datasets and across time; so for spine construction purposes there has to be a central linkage procedure with full information at least for a limited time. For some purposes (eg sampling, evaluating nonresponse) personal identifiers are not needed so a pseudonymised output dataset could be constructed with personal identifiers dropped. However, location identifiers are needed, and these mean that the disclosure risk will remain high even in a pseudonymised dataset, for the kinds of purposes suggested for investigation in the Longitudinal Studies Review, like evaluating nonresponse and coverage biases.

5.2 Using a spine

The key purpose of a spine is as infrastructure to support the linking of data. Most existing research is not based on the spine information directly, but needs to define a single, primary data source which is the basis of a population of interest. Variables can be added to this source from other sources by linkage through the spine. The primary source need not be an administrative data source, and there are already examples of surveys being used in this way. The Swedish Gender and Generations Survey (https://www.ggp-i.org/data/methodology/) is a two-wave longitudinal study which combines data collected in wave one from a telephone/online survey in 2012/2013 to data available from the Swedish Population Register. Follow up in the second wave was based entirely on data available within the Population Register. Outcomes are currently available to 2016, but in theory individuals can be followed up for longer periods of time within Population Registers and life course dynamics (such as leaving and returning to the parental home) can be examined using a 'linked lives' perspective since Population Registers link data from biologically related individuals (Kleinepier et al., 2017).

To support any particular study there is a need to define the population of interest, and this is already more complicated for longitudinal surveys than for cross-sectional surveys because of the temporal element in the definition (Smith *et al*. 2009, Lynn in Goldstein *et al*. 2015). To use a spine as a basis for assessing (for example) the non-response bias or differential coverage of different population subgroups, we would first need to define which units on the (maximal) spine are part of the population of interest. This is often not straightforward; people do not deregister from administrative systems when emigrating, and there may be lags between events occurring and records on the spine being updated, so that people appear to be in different areas than they really are. One approach would be to get indicative data quickly (perhaps to allow some form of responsive design to compensate for deficiencies in recruitment), with a more detailed later evaluation when most of the lags have worked out, to give detailed information on which to base adjustments to estimates.

There is also the possibility that people will be duplicated, or missed within the administrative data. If some procedure (for example a survey, or a search for duplicates of specific records) is used as the basis of an assessment, then it will generally be possible to use the information obtained to come to a judgement about which of a set of duplicate records is correct at a given time. However, records which are undercovered are by definition not on the spine, and therefore cannot be sampled.

5.3 Development of procedures for an administrative data spine

Davis-Kean et al. (2016) have made the case for use of administrative data to support and enhance longitudinal data collection as an infrastructure for a wide range of analyses. The details of how such this would be put together and analysed still need to be worked out. A single administrative source is more easily understood and interpreted and does not require linkage for its construction. But it may not contain enough information on important subpopulations to allow for the envisaged adjustments of supplementation of a survey collection.

There are also interesting questions in how to present and analyse a dataset containing complete (or almost complete) administrative data for selected units (eg addresses), together with data only for participating units on important analytical variables. The administrative data could be used in the calculation of non-response adjustment weights (Särndal & Lundström 1999, Haziza & Lesage 2016), but calibration to more than a few variables may increase variances by increasing variability in the weights. There is an outstanding issue between the desirability of bespoke weights for particular analyses, or general purpose weights for consistency, and some question over whether researchers make informed decisions to use or not the weights which are already provided. An alternative approach would be to use a model to produce the best estimates using all of the available information (as in Skinner & Coker (1996) who use maximum likelihood estimation to take advantage of the wider information available in the larger dataset); application of fractional imputation is another possible approach (Yang & Kim in press). These approaches require considerable methodological sophistication, however, and suggest a need for further training in advanced quantitative methods.

A simpler approach would be to use these data as a quality evaluation for the achieved sample, as a standard reference for researchers to cite. While likely to be well-used, this approach however does not adjust the standard analyses for the characteristics of the achieved sample, and so may not make best use of the available information.

5.4 Triggers for data collection

A data spine can also behave as an early warning system for events which are likely to be of particular interest in the life course, and prompt an additional contact. For example a period of unemployment signalled in an administrative source might be followed up by some additional questions on changes to financial behaviour or on job search strategies and outcomes.

6. Data linkage

6.1 Data linkage methods

A variety of methods have been proposed for data linkage in the absence of unique identifiers, which is a frequent occurrence in administrative data linkage in the UK. The traditional approach is based on the Fellegi & Sunter (1969) paradigm, and there are various approaches to implement this in practice, with the EM algorithm a popular approach (see Herzog *et al*. 2007, Chapter 9, for an overview). This algorithm assigns candidate pairs to be links or non-links, or a third category where there is insufficient information to make an automatic choice, and where clerical review is needed. The algorithm is computationally intensive, and treats different matching variables as if they are independent, when in fact there may be varying amounts of discriminatory information in each.

Recently, Goldstein *et al*. (2017) have developed an alternative approach which is less computationally intensive, and which derives for each variable to reflect the amount of discriminatory information they contain, in a similar way to multivariate analysis methods, thus avoiding the independence assumption. More experience with the impact of Goldstein *et al*. (2017)'s approach in practical situations is needed.

6.2 Quality of linkage

Assessing the quality of automated data linkage generally requires clerical review of a sample of links. Where the quality of the linkage needs to be high, the required sample size may be large to provide estimates of quality measures of sufficient accuracy (ONS 2013). Boyd *et al*. (2016) have suggested a simple sampling and estimation procedure, and further experience with sampling strategies (especially for false negatives, which are comparatively very rare) are needed.

References

Black, A (2016). *The IDI prototype spine's creation and coverage*. (Statistics New Zealand Working Paper No 16–03). Retrieved from www.stats.govt.nz.

Böhning, D., van der Heijden, P.G.M. & Bunge, J. (eds) *Capture-recapture methods for the social and medical sciences*. CRC Press: Boca Raton.

Boyd, J., Guiver, T., Randall, S., Ferrante, A., Semmens, J., Anderson, P. & Dickinson, T. (2016) A simple sampling method for estimating the accuracy of large scale record linkage projects. *Methods of Information in Medicine* **55** 276-283.

Callegaro, M. (2008) Seam effects in longitudinal surveys. *Journal of Official Statistics* **24** 387-409.

Colledge, M.J. (1995) Frames and business registers: an overview. Pp. 21-47 in B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott (eds), *Business survey methods.* New York: Wiley.

Connelly, R., Playford, C.J., Gayle, V. & Dibben, C. (2016) The role of administrative data in the big data revolution in social science research. *Social Science Research* **59** 1-12.

Courgeau, D. & Najim, J. (1996) Interval-censored event history analysis. *Population: an English selection* **8** 191-207.

Davis-Kean, P., Chambers, R.L., Davidson, L.L., Kleinert, C., Ren, Q. & Tang, S. (2017) *Longitudinal Studies Strategic Review*. Report to the Economic and Social Research Council.

Dezateux, C., Knowles, R., Brocklehurst, P., Elias, P., Burgess, S., Colson, D., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Kiernan, K., Leon, D., Reay, D., Sera, F., Vignoles, A. & Walton, S. (2016) *Life Study Scientific Protocol*. (Life Study Working Papers). Life Course Epidemiology and Biostatistics/ UCL Institute of Child Health: London, UK. http://discovery.ucl.ac.uk/1485668/.

Dezateux, C., Colson, D., Brocklehurst, P. & Elias, P. (2016) *Life after Life Study*. http://discovery.ucl.ac.uk/1485681/.

Dunne, J. (2015) The Irish Statistical System and the emerging Census opportunity. *Statistical Journal of the IAOS* **31** 391–400. https://doi.org/10.3233/SJI-150915.

Fellegi, I.P. & Sunter, A.B. (1969) A theory for record linkage. *Journal of the American Statistical Association* **64** 1183-1210.

Goldstein, H., Lynn, P., Muniz-Terrera, G., Hardy, R., O'Muircheartaigh, C., Skinner, C.J. & Lehtonen, R. (2015) Population sampling in longitudinal suverys. *Longitudinal and Life Course Studies* **6** 447-475. https://doi.org/10.14301/llcs.v6i4.345.

Hand, D. J. (2018) Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society, Series A* **181** 555-605.

Haziza, D., & Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics* **32** 129-145.

Herzog, T.N., Scheuren, F.J. & Winkler, W.E. (2007) *Data quality and record linkage techniques*. Springer Science & Business Media: New York.

HM Treasury (2014) Use of Administrative Sources for Statistical Purposes, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/361919/admin_sources_Oct_14.pdf.

Hurren, H., Stewart, A. & Dennison, S. (2017). New methods to address old challenges: the use of administrative data for longitudinal replication studies of child maltreatment. *International Journal of Environmental Research and Public Health* **14** 1066, https://doi.org/10.3390/ijerph14091066.

Kleinepier, T., Berrington, A. & Stoeldraijer, L. (2017) Ethnic differences in returning home: explanations from a life course perspective. *Journal of Marriage and Family* **79** 1023-1040.

Lessof, C. (2009). Ethical issues in longitudinal surveys. Pp 35-54 in P. Lynn (ed) *Methodology of longitudinal surveys*. Wiley: Chichester.

Lundström, S., & Särndal, C. E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics* **15** 305-327.

Maasing, E., Tiit, E-M & Vähi, M. (2017) Residency index - a tool for measuring the population size. Acta et Commentationes Universitatis Tartuensis de Mathematica Volume 21, Number 1, acutm.math.ut.ee/index.php/acutm/article/view/ACUTM.2017.21.09/74.

Meyer, B.D. & Mittag, N. (2019) Using linked survey and administrative data to better measure income: implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics*, 11, 176-204. DOI: 10.1257/app.20170478.

Mostafa, T. (2016). Variation within households in consent to link survey data to administrative records: evidence from the UK Millennium Cohort Study. *International Journal of Social Research Methodology*, 19(3), 355-375. https://www.tandfonline.com/doi/abs/10.1080/13645579.2015.1019264

ONS (2013) An assessment of the quality of the matching between the 2011 Census and the Census Coverage Survey. https://webarchive.nationalarchives.gov.uk/20160108085304/http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/census-coverage-survey--ccs-/matching-quality-report-ns.pdf

ONS (2016) Statement of administrative sources. Available from https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/admindatasources

Randall, S.M., Ferrante, A.M., Boyd, J.H. & Semmens, J.B. (2013) The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making* **13** 64.

Ruiz, M., Benzeval, M. & Kumari, M. (2017) A guide to the biomarker data in the CLOSER studies. CLOSER: London. https://www.closer.ac.uk/wp-content/uploads/A-guide-to-the-biomarker-data-in-the-CLOSER-studies-FINAL.compressed.pdf.

Sakshaug, J.W. & Antoni, M. (2017) Errors in linking survey and administrative data. Pp 557-573 in P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker & B.T. West (eds) *Total survey error in practice*. John Wiley & Sons, Inc: Hoboken.

Sakshaug, J.W. & Kreuter, F. (2012) Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods* **6** 113-122.

Sexton, A., Shepherd, E., Duke-Williams, O., & Eveleigh, A. (2017). A balance of trust in the use of government administrative data. *Archival Science* **17** 305-330.

Skinner, C. J., & Coker, O. (1996). Regression analysis for complex survey data with missing values of a covariate. *Journal of the Royal Statistical Society: Series A* **159** 265-274.

Smith, P., Lynn, P. & Elliot, D. (2009) Sample design for longitudinal surveys. Pp 21-33 in P. Lynn (ed) *Methodology of longitudinal surveys*. Wiley: Chichester.

Tourangeau, R. (2018) The survey response process from a cognitive viewpoint. *Quality Assurance in Education* **26** 169-181, https://doi.org/10.1108/QAE-06-2017-0034.

Van Delden, A., Pannekoek, J., Banning, R. & Boer, A.D. (2016) Analysing correspondence between administrative and survey data. *Statistical Journal of the IAOS* **32** 569-584.

Van Delden, A., Scholtus, S. & de Wolf, P.-P. (2011) *The case of the missing tax data*. Statistics Netherlands Discussion Paper 201115. Statistics Netherlands: The Hague/Heerlen. https://www.cbs.nl/-/media/imported/documents/2011/12/2011-x10-15.pdf.

Van den Brakel, J.A., Smith, P.A., Elliott, D., Krieg, S., Schmid, T. & Tzavidis, N. (2020, forthcoming) Assessing discontinuities and rotation group bias in rotating panel designs. Chapter 16 in Lynn, P. (ed) *Advances in Longitudinal Survey Methodology*. Wiley: Chichester.

Van der Heijden, P.G.M., Smith, P.A., Cruyff, M. & Bakker, B. (2018) An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *Journal of Official Statistics* **34** 239–263. doi: dx.doi.org/10.1515/JOS-2018-0011.

Warren, J.R. & Halpern-Manners, A. (2012) Panel conditioning in longitudinal social science surveys. S*ociological Methods & Research* **41** 491-534.

Watson, N. & Lynn, P. (2020, forthcoming) Refreshment sampling for longitudinal surveys. In P. Lynn (ed) *Advances in Longitudinal Survey Methodology*. Wiley: Chichester.

Yang, S. & Kim, J.K. (in press) A semiparametric inference to regression analysis with missing covariates in survey data. *Statistica Sinica*. http://www3.stat.sinica.edu.tw/ss_newpaper/SS-14-174_na.pdf.

Zhang, L.-C. (2011) A unit-error theory for register-based household statistics. *Journal of Official Statistics* **27** 415-432.

Zhang, L-C. (2015) On modelling register coverage errors. *Journal of Official Statistics* 31 381-396. https://doi.org/10.1515/jos-2015-0023.

Zhang, L.-C. & Fosen, J. (2012) A modeling approach for uncertainty assessment of register-based small area statistics. *Journal of the Indian Society of Agricultural Statistics* **66** 91-104.

.