International Journal of Social Research Methodology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713737293

The Feasibility of Linking Household Survey and Administrative Record Data: New Evidence for Britain

Stephen P. Jenkins; Peter Lynn; Annette Jäckle; Emanuela Sala

First Published on: 06 November 2007
To cite this Article: Jenkins, Stephen P., Lynn, Peter, Jäckle, Annette and Sala, Emanuela (2007) 'The Feasibility of Linking Household Survey and Administrative Record Data: New Evidence for Britain', International Journal of Social Research Methodology, 11:1, 29 - 43
To link to this article: DOI: 10.1080/13645570701401602
URL: http://dx.doi.org/10.1080/13645570701401602

Int. J. Social Research Methodology
Vol. 11, No. 1, February 2008, pp. 29–43

The Feasibility of Linking Household Survey and Administrative Record Data: New Evidence for Britain

Stephen P. Jenkins, Peter Lynn, Annette Jäckle & Emanuela Sala

Received 25 February 2006; Accepted 11 April 2007

Linkage of household survey responses with administrative data is increasingly on the agenda. Unique individual identifiers have clear benefits for making linkages but are also subject to problems of survey item non-response and measurement error. Our experimental study that linked survey responses to UK government agency records on benefits and tax credits elucidates this trade-off. We compare five linkage criteria: one based on a respondent-supplied National Insurance Number (NINO) and the other four using different combinations of sex, name, address and date of birth. As many linkages were made using non-NINO-based matches as were made using matches on NINO and the former were also relatively accurate when assessed in terms of false-positive and false-negative linkage rates. The potential returns from hierarchical and pooled matching are also examined.

Introduction

Although linkage between household survey responses and administrative records is rare in Britain (Plewis, Smith, Wright, & Cullis, 2001), it is increasingly on the agenda. For example, the English Longitudinal Study of Ageing is supplementing survey data with information about respondents' National Insurance contribution histories, benefit and tax credit records held by government agencies and information

Stephen P. Jenkins and Peter Lynn are Professors at the Institute for Social and Economic Research (ISER), University of Essex. Annette Jäckle and Emanuela Sala are Senior Research Officers at ISER. As part of the UK Economic and Social Research Council (ESRC) Research Methods Programme, the authors studied methods to improve survey measurement of income and employment. Correspondence to: Stephen P. Jenkins, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. Tel.: +44 (01206) 873374; Fax: +44 (01206) 873151; Email: stephenj@essex.ac.uk

ISSN 1364–5579 (print)/ISSN 1464–5300 (online)
DOI: 10.1080/13645570701401602

from hospital episode statistics and from mortality and cancer registration records (Marmot, Banks, Blundell, Lessof, & Nazroo, 2003). The Office for National Statistics (ONS) and the Department for Work and Pensions (DWP) have a pilot project investigating the feasibility of linking administrative record data on benefits to working-age respondents to the Labour Force Survey. The Millennium Cohort Study is including data obtained from hospital episode statistics and birth registration records, and plans to include school records in later sweeps (Dex & Joshi, 2004). In general, record linkage has several attractions for household survey producers and users: it may help diminish respondent burden, additional information may be collected, and measurement error may be reduced. For further discussion of the possibilities provided by linked administrative record data, see Calderwood and Lessof (2006) and Jones and Elias (2006).

In this article, we provide evidence about the feasibility of such linkages and an important related issue: the choice of variables to be used to implement the link between respondents in the survey and records in the administrative source. Our analysis is based on an experimental study that linked UK government agency records on benefits and tax credits to household survey respondents. We highlight the advantages and disadvantages of using a self-reported National Insurance Number (NINO) as a linkage key. Although their primary purpose relates to the administration of the National Insurance system, NINOs are widely used because, along with National Health Service numbers, they are unique identifiers available to all adults in the UK.

We compare the NINO-based linkage criterion with four other criteria for linking survey respondents to administrative records on receipt of benefits and tax credits held by the Information Centre of the DWP. Distinctive features of our work include the examination of the relative performance of the five matching criteria (and their combination) in terms of the numbers of matches made and their relative accuracy. Although our analysis is based on a British household survey (the 'Improving Survey Measurement of Income and Employment', i.e. ISMIE survey), the issues addressed are of wider relevance. The match criteria that we use are similar to those that are available in most household surveys in most countries.

Although the methodological issues of data linkage are relatively well known, there is little experience in Britain of linkage of administrative records to household survey data. The examples cited in the survey by Jabine and Scheuren (1986) refer to the USA. A review of issues for the UK (Gill, 2001) focuses on linkage between different types of administrative sources, especially medical records. The two previous UK studies that have linked survey and administrative data did not compare matching strategies: see Noble and Daly (1996) studying Disability Living Allowance claimants, and the DWP (2003) study linking eligible non-recipients of the Minimum Income Guarantee (MIG) who were respondents to the 2000/2001 Family Resources Survey with administrative records on benefits. Moreover, both these studies focused on narrowly defined subgroups of the population: disabled persons and low-income pensioners. Our study uses a more widely defined population sample.

In the rest of this Introduction, we briefly review some methodological issues concerning matching (see Jabine & Scheuren, 1986 and Gill, 2001 for detailed reviews).

In the next section, we describe the ISMIE survey and methods of linkage with the DWP data. We compare the linkage rates of the various match criteria, and assess their relative accuracy, in the following two sections. Our investigation of the sources of mismatch and measurement error provides guidance about how to refine matching criteria in future linkage exercises using survey data. The final section contains a summary and conclusions.

Methodological Issues Concerning Linkage Criteria

The advantages of matching using a personal identifier such as a NINO are clear: a NINO is unique to each individual and virtually all adult Britons have one, and so it has a very high discriminatory power. And, once issued, it does not change. There are, however, potential disadvantages to NINO-based matches when a NINO is derived from a household survey. There is possible *item non-response*: respondents may not be willing to provide a NINO or they may not know what their NINO is. This lowers the number of linkages that are possible to make. There is also potential *measurement error*: respondents may report NINOs with error or interviewers may transcribe them incorrectly. As with the US Social Security Number, the NINO does not contain a check digit (a single digit computed from the other digits in the number). One can only check the basic format: two letters followed by six digits.

Instead of, or as well as, linking records using a NINO, one can use variables that are collected as part of the survey and which also appear in the administrative database. The advantage of this strategy is that the variables are already available, and there is no additional respondent burden. The disadvantages of using these variables are as follows. First, a match may not be unique, even if several variables are used in combination. Information about the sex of a respondent may have high accuracy but it also has low discriminatory power (there are only two sexes).

Second, there is potential for mismatch because the survey and the administrative source may record the same information differently. This may reflect measurement error in either or both of the sources (e.g. a misspelt name or incorrect postcode), or different recording conventions (e.g. administrative records may record a respondent's first name as William, and the survey record it as Bill). Pre-processing of name and address information is a commonly used method of reducing the effects of these factors.

In addition, what is apparently the same piece of information in the two sources may refer to different dates if the survey data were collected after the date to which the administrative data refers, or vice versa. For example, residential mobility may lead to mismatch by address and marriage or divorce could lead to a mismatch by name if there is a change in family name. The longer the interval between the two reference dates, the greater these problems are likely to be.

In sum, the choice of matching variables for linkage between survey and administrative data requires information about the numbers of matches made using different match criteria, and about their relative accuracy. Our experimental study provides new evidence about this for the UK.

Linkage of Data from the ISMIE Survey and DWP Administrative Records

The ISMIE Survey

The household survey data were derived from the ISMIE survey, a follow up to the 2001 wave of the BHPS-ECHP panel. This panel was derived from a random sample of private households, the UK component of the European Community Household Panel Survey (ECHP-UK). This began in 1994, with annual interviews thereafter. Following the major reorganisation in ECHP design in the mid-1990s, a sub-sample was drawn from the ECHP-UK and surveyed jointly with the primary samples of the British Household Panel Survey (BHPS) from 1997 onwards. Although the original sub-sample aimed to focus on low income households, the realised sample contained a notable number of households with middle-range incomes, and some with high incomes. Funding for the BHPS-ECHP sub-sample expired in 2001, and hence the previously regular cycle of interviewing stopped. However, with Economic and Social Research Council (ESRC) funding, we had the opportunity to interview respondents once more for purely methodological purposes.

The ISMIE fieldwork took place in Spring 2003. Interviews were sought with all BHPS-ECHP panel members who had responded in survey year 2001, i.e. 1167 individuals aged 16+ in 785 households. Eligible movers were followed to their new address. The achieved sample with complete interviews was 1033 adults, i.e. 89% of the eligible sample. The ISMIE questionnaire was the same as that given to the main BHPS sample in Autumn 2002, except that some modules were added for the purposes of the methodological work, and some others (e.g. about health) were excluded in order to minimize total respondent burden and to economise on survey costs. For further details of the ISMIE survey, see Jäckle, Jenkins, Lynn, and Sala (2004).

At the end of the individual interview, the interviewer read a preamble stating that additional analysis was being undertaken that year especially to assess the quality of data collected in the survey. Then respondents were asked whether they were happy to give us permission to link their answers with the administrative records held by the DWP and Inland Revenue about their benefits and tax credits (but not about their income tax). Everyone who gave consent was asked for their National Insurance Number, and requested to consult a payslip or other records such as a pension or benefit book or NINO card. (Whether they did or not was recorded.) The Computer-Assisted Personal Interviewing (CAPI) script checked that the NINO provided was of the correct format. Data linkages were sought for all consenting respondents, regardless of whether they had reported receipt of benefits. As a significant minority of respondents had never received benefits, and so were not cases on the DWP database, the maximum possible linkage rate was less than 100%. We return to this issue below.

The DWP Administrative Data

Our data linkages were to information held in the DWP's '100% Generalized Matching Service' Primary Data file. This file contains a record for each person who is currently receiving, or has ever received, any one of 15 benefits. These include Child Benefit,

Housing Benefit, Working Families Tax Credit, several types of disability benefit, Income Support, Job Seeker's Allowance and the state retirement pension.

Each record contains personal details derived from information collected when a benefit claim was made, and is updated when new scans of benefit receipt databases indicate that they have changed. The personal details include NINO, title (Mr, Miss, Ms and Mrs; and hence sex), date of birth (day, month, year), first name, family name, address, and postcode. Each of these variables was potentially available from the ISMIE survey too, and they were the basis of our linkage experiment.

For the ISMIE project, the Primary Data file was accessed in the week beginning 13 October 2003, i.e. several months after the survey interview. Information for each recipient about dates of receipt and amounts paid is held by the DWP in separate files, each linked to the Primary File using the individual's NINO as the key. (The information in these separate files is obtained from regular 'scans', that is 100% data extracts of all current claims, taken as a snapshot at a particular date. Income Support and Job Seekers Allowance data are extracted every two weeks; Child Benefit, Disability Living Allowance, Attendance Allowance, Industrial Injuries Disablement Benefit, Invalid Care Allowance and Tax Credit data are extracted every four weeks; Retirement Pension, Widows Benefit, Bereavement Benefit, Severe Disablement Allowance and Incapacity Benefit data are extracted every six weeks.) Histories of benefit and Tax Credit receipt were obtained, covering the period 1999 to 2003.

The Match Criteria Used for Linkage

Five independent matching criteria were used to link consenting ISMIE survey respondents to the DWP Primary Data. (We specified the criteria; the linkages per se were undertaken by DWP staff.) The match criteria were characterized by the following sets of variables:

Criterion 1: NINO
Criterion 2: Sex, date of birth, postcode
Criterion 3: Sex, date of birth, forename, family name
Criterion 4: Sex, postcode, forename, family name
Criterion 5: Sex, forename, family name, address line 1

Sex was either male or female. Date of birth had day, month and year fields. UK postcodes have two parts. The first, the 'outward code', is one or two letters denoting the Area followed by one or two digits, denoting the District. The second part, the 'inward code', is a digit followed by two letters (the Unit). There are 9473 postal sectors (defined by outward code plus inward code digit) in Britain, with an average of about 2530 addresses per sector (Lynn & Lievesley, 1991). An example of 'address line 1' is '12 Errol Street'.

Because the five linkage exercises were undertaken independently, we could also combine the results to simulate the effects of using various hierarchical match criteria whereby a second criterion is applied to respondents who are not successfully matched with a first criterion, etc. We focused on two criteria involving NINOs (criterion 1

followed by criterion 2 applied to cases not matched on criterion 1, and vice versa), and two criteria based on non-NINO matching (criterion 2 followed by criterion 3, and vice versa). The latter two criteria are similar to the criteria used by the DWP (2003) study. Finally, we also considered the effects of pooling the results of all five linkages. Contrasting the linkages made using different criteria highlights the impacts of each of the different matching variables, the NINO in particular.

Although this analysis is the first of its kind for Britain, it was not a full-fledged linkage implementation study. For example, we used only exact (deterministic) matching, whereas one could also use probabilistic matching (Gill, 2001) and more extensive preprocessing of data. We used survey variables verbatim, apart from the cleaning and formatting already implemented as part of routine panel maintenance and follow-up. The DWP variables were also used verbatim except that addresses and postcodes had already been processed into a consistent format using proprietary QuickAddress Software (QAS™), an option not possible with the survey data given the resources available. The research reported here was a one-shot experimental study piggybacking on a larger survey methodological project.

Linkage Rates

Before undertaking record linkages for ISMIE respondents, we had to gain their informed consent. Consent rates were relatively high. About 78% of the sample provided consent, with no differences in the rates for men and women. Some 88.7% of consenting ISMIE respondents supplied a NINO, with little difference in the fraction for men (87.4%) and women (89.4%). Put another way, 68.8% of the ISMIE sample provided both consent and a NINO. For a detailed analysis of ISMIE respondents' consent and NINO supply propensities, see Jenkins, Cappellari, Lynn, Jäckle, & Sala (2006). For discussion of informed consent and related ethical issues relating to administrative record linkage, see ONS (2004) and Lessof (2006).

The main reason stated for not supplying a NINO was that the respondent did not know it (9.9% of the sample), rather than a refusal to provide it (1.5%). The data also suggest that reported NINOs are likely to be reliable. Among respondents who supplied a NINO, just over two-thirds (67.4%) referred to a payslip or other document, and 30.8% supplied the number from memory but were confident that the number was correct. Only 1.8% stated that they were not sure about the NINO supplied. The rate of consultation of documents to check the NINO was markedly higher among respondents aged 50+ (81.2%) than among respondents aged less than 50 (54.3%). This suggests that older people were less confident in remembering their NINOs or simply that pension books were more readily available than payslips.

We now examine linkage success rates for the five match criteria. Recall that there are two potential reasons for a linkage not being made. Either the relevant ISMIE respondent had never received one of the benefits or tax credits for which the DWP database has information (a 'true non-match'), or the respondent had received one of the benefits or tax credits but could not be linked using the five match criteria (a 'false non-match'). We estimate that the true non-match rate is about one quarter, because

71% of the ISMIE respondents reported receiving at least one of the relevant benefits or tax credits at least one annual interview between 1999 and 2003.

The 'pooled' linkage rate, i.e. counting all matches on at least one criterion, was 57.3%, which is about 18 percentage points lower than the rate expected if there were no false non-matches. This suggests that there are false non-matches, but it is difficult to assess their prevalence further because there are no comparable matching exercises against which to benchmark our results. The linkage rate for matches between respondents to the US Health and Retirement Study (HRS) and earnings records held by the Social Security Administration (SSA), made using Social Security Numbers, was 75% (Olson, 1999). However, this rate is not comparable with the overall ISMIE one, or the NINO-based rate discussed below, as the expected true non-match rate is lower than in our study. In the HRS the expected true non-match rate is near zero: virtually all US adults aged 50+ have had some labour earnings during their working life and hence an SSA record. In the DWP (2003) study that matched low-income pensioner respondents from the Family Resources Survey with DWP records, the expected true non-match rate was also negligible, because virtually all of the respondents would have been receiving a retirement pension or a winter fuel payment. The actual match rate was 96% (2003, p. 55).

The linkage rates for each of the various independent and hierarchical criteria are shown in Table 1. (These are the raw linkage rates, and potentially include mismatches, which are discussed further below.) Among the independent matching exercises, the greatest linkage rate was for matching based on sex, date of birth and postcode (criterion 2), followed closely by matching based on NINO (criterion 1) and sex, date of birth, forename and family name (criterion 3). The rates were 49.7%, 48.2% and 47.9%, respectively, when expressed as a fraction of the ISMIE sample size (Table 1, column 1), or 64.0%, 62.1% and 61.7%, when expressed as a fraction of the number of consenting respondents (column 2). Matching by criterion 4, and especially by criterion 5, led to noticeably worse linkage rates, suggesting that date of birth and sex together have relatively high discriminatory power and/or that address and name data are subject to more variation in how they are recorded. We return to this issue below. Almost three quarters of all consenting respondents were matched by at least one criterion ('pooled' matching).

The high potential return to hierarchical matching is shown in the lower panel of Table 1. Employing two criteria in combination identified a significant number of additional matches, for both NINO-based and non-NINO-based hierarchical matches. For hierarchical matches based on criteria 1 and 2, and on criteria 2 and 3, the linkage rate was only about one percentage point below the rate achieved by pooling independent matches on any one of the five criteria.

Many of the differences between linkage rates for the NINO-based match and for matches based on sex and date of birth (criteria 2 and 3) arose because of NINO item non-response: see columns 3 and 4 of Table 1. Interestingly, the linkage rates for criteria 2–5 were all lower for respondents who did not supply a NINO than for those who did. This might be indicative of a general tendency to supply lower quality data, or it may be that respondents who receive benefits are more likely to supply a NINO. Among respondents who supplied a NINO, the linkage rate when matching by NINO was 71%.

Table 1 Record Linkage Rates (%) for ISMIE Respondents

Criterion and matching variables	ISMIE sample (1)	All who gave consent to data linkage (2)	Supplied NINO (3)	Did not supply NINO (4)
Independent matching				
1. NINO	48.2	62.1	70.0	–
2. Sex, date of birth, postcode	49.7	64.0	64.3	61.5
3. Sex, date of birth, forename, family name	47.9	61.7	62.6	55.0
4. Sex, postcode, forename, family name	41.7	53.7	54.4	48.4
5. Sex, forename, family name, address line 1	33.7	43.4	44.3	36.3
Pooled matching: at least one of the above	57.3	73.8	74.5	68.1
Hierarchical matching				
1 followed by 2, or 2 followed by 1	56.4	72.6	74.1	61.5
2 followed by 3, or 3 followed by 2	56.1	72.1	72.7	68.1
N	1033	802	711	91
(as % of all who gave consent)		(100)	(88.7)	(11.3)

Note: The table includes potential mismatches (see section 'Linkage Rates' in the text). NINO, National Insurance Number; ISMIE, Improving Survey Measurement of Income and Employment.

Table 1 might also be interpreted as saying that matching by non-NINO criteria is a potential strategy for record linkage in the future, given that securing a NINO from every survey respondent is problematic. The veracity of this conclusion depends on the accuracy of the various linkages. Before turning to this issue, we consider the overlaps between the sets of respondents for whom linkages were made.

Table 2 lists the combinations of linkage outcomes from the five independent matching exercises. Of the respondents who gave linkage consent, 26% were not linked by any of the five independent criteria, 4% were linked by one criterion, 15% by two criteria, 4% by three criteria, 15% by four criteria and 36% were linked by all five. The degree of overlap between the respondents identified by even the most successful match criteria is perhaps surprisingly small. For example, 155 respondents (19% of all consenting respondents) were matched either by criterion 1 or by criterion 2, but not by both. At the same time, this highlights again the potential return to hierarchical or pooled matching procedures.

Table 2 also confirms the impression that criteria 4 and 5 add very little to the other three criteria. Pooled matching using only criteria 1 to 3 produces exactly the same result as pooled matching using all five criteria, as there are no respondents who match only on one or both of criteria 4 or 5.

Linkage Accuracy

The accuracy of linkage by a particular criterion may be assessed along two dimensions. First, one wants to minimize the proportion of actual matches that are erroneous matches. This is the *false-positive rate*, calculated for criterion *m* as the number of mismatches by *m* divided by the total number of matches by *m*. Second, one also wishes

Table 2 Linkage Outcomes among Consenting ISMIE Respondents

	All who gave consent to data linkage		All who gave consent and supplied a NINO	
Linkage outcomes*	Frequency	Percentage	Frequency	Percentage
00000	210	26.2	181	25.5
00100	7	0.9	3	0.4
00101	2	0.3	0	0
01000	16	2.0	4	0.6
01110	20	2.5	7	1.0
01111	49	6.1	18	2.5
10000	11	1.4	11	1.6
10010	1	0.1	1	0.1
10011	1	0.1	1	0.1
10100	47	5.9	47	6.6
10101	10	1.3	10	1.4
11000	68	8.5	68	9.6
11110	74	9.2	74	10.4
11111	286	35.7	286	40.2
All	802	100.0	711	100.0

* Outcomes for criteria 1, 2, 3, 4, and 5 (in that order), with '0' meaning not matched, and '1' meaning matched. For example '10010' means respondent matched by criteria 1 and 4, but not by 2, 3 or 5.
Note: The match criteria are defined in the text and summarised in Table 1. The table includes potential mismatches (see section 'Linkage Rates' in the text).

to minimize the proportion of non-matches that are erroneous. This *false-negative rate* is calculated for criterion *m* as the fraction of non-matches by *m* that were genuine matches according to criteria other than *m*. (The rate is defined relative to a specific set of criteria.) For a given linkage rate, one match criterion is unambiguously better than another if the first has a lower false-positive rate and a lower false-negative rate than the second. If this is not the case, unambiguous rankings of match accuracy involve additional judgements about the appropriate trade-off between the risks associated with false positives and those associated with false negatives.

We estimated false-positive and false-negative rates by pooling information from the five independent matching exercises. For example, for NINO matches, the false-positive rate was derived from information on cases with match patterns of form '1xxxx' in Table 2, and the false-negative rate was derived from information on cases with match patterns of form '0xxxx' (where 'x' refers to '0' or a '1'). Estimates were calculated for criteria 1–3 (but not for criteria 4 and 5 given their relatively low match rates), and for the hierarchical and pooled criteria discussed earlier. When calculating false-negative rates, the appropriate treatment of the 210 cases not matched on any criterion (pattern '00000' in Table 2) is a moot point: as explained earlier, a majority of these respondents were likely to be true non-matches (non-recipients of benefits). We report estimates of false-negative rates based on the assumption that all these individuals were non-recipients of benefits. Supposing instead that they were all benefit recipients increased the magnitude of every estimate but did not change the ordering of the criteria by false-negative rate.

We assumed that matches made by three or more of the five independent matching criteria were genuine matches (except in one NINO-related situation discussed shortly), and inspected listings of information about all remaining cases to assess whether an actual match (or non-match) was true or false. Although, this introduced an element of researcher judgement, assessment was almost always clear cut in practice. For example, when the survey and DWP postcodes differed, they usually did so by only one or two characters, and it was clear from the name, address and birth date information, that the correct person had been identified according to one or more other criteria. Address information is discussed further below.

The exceptional NINO-related situation was when the matching process led to two different individuals in the DWP Primary Data (with two different NINOs) being associated with a single respondent in the ISMIE survey. This arose with 14 respondents (13 with match pattern '11111' and one with '11000'). We could determine that, in eight cases, the NINO from the survey was incorrect and hence there was a mismatch by criterion 1 but a genuine match by other criteria. In three cases, there was a mismatch by criterion 3, and in one case, mismatch by criterion 5.

The estimates of the false-positive and false-negative linkage rates are shown in Table 3. In several of the table cells, a range has been reported rather than a single estimate. In each of these cases, estimation involved comparisons of address information. Visual inspection could not resolve with certainty whether there was a genuine match or genuine mismatch, since addresses could legitimately differ between the survey and DWP databases because of residential mobility.

The match pattern '10100', i.e. a match by NINO and also by sex, date of birth, forename and family name, illustrates problems arising with address information. The 47 respondents had an 'address line 1' that differed between the survey and the DWP file.

Table 3 Estimates of Linkage Accuracy

	False-positive rate		False-negative rate	
Matching method	%	(*N*)	%	(*N*)
Independent matching				
1. NINO	2.2–11.6	(498)	30.9	(304)
2. Sex, date of birth, postcode	0	(513)	23.9–27.3	(289)
3. Sex, date of birth, forename, family name	0–10.9	(495)	30.6	(307)
Hierarchical matching				
1 followed by 2	1.9–9.9	(583)	4.1	(219)
2 followed by 1	0.5–8.6	(583)	4.1	(219)
2 followed by 3	0–8.1	(579)	4.7	(213)
3 followed by 2	0–9.3	(579)	4.7	(213)
Pooled matching				
Match by at least one of 1–5	0–8.6	(592)	0	(210)

Notes: Independent, hierarchical and pooled matching defined in the text. False-positive rate for criterion *m* = percentage of matches by *m* that were mismatches according to criteria other than *m*. False-negative rate for criterion *m* = percentage of non-matches by *m* that were genuine matches according to criteria other than *m*. Estimates of false-negative rates assume that all 210 cases with match pattern '00000' were not benefit recipients (see text). *N* refers to the number in the denominator of the relevant rate calculation.

However, inspection revealed that three cases had virtually identical address line 1 and postcode (so the errors probably reflected transcription errors), 23 were in the same postal Area and District (i.e. had the same outward code), 15 were in the same postal Area and there were six other cases. We believe that most of the respondents were identified correctly since most residential mobility in Britain is short distance (Böheim & Taylor, 2000, Table 1). Readers sharing our belief should take the estimates of false-positive rates as lying towards the lower end of the range shown, and vice versa for the false-negative rate.

The lowest false-positive rate among the independent matching criteria was for matches by sex, date of birth and postcode (criterion 2): a remarkable 0%. The rates for NINO matches and criterion 3 were several percentage points higher depending on how the information about addresses was treated. The rate in the former case was at least 2.2%, highlighting the fact that NINOs derived from surveys are subject to measurement error.

NINO measurement error is illustrated by the data for the 32 respondents who supplied a NINO and for whom there was a match on one or more criteria other than the NINO. In 10 cases, the first two letters of the NINO were in error; for example the letters 'M' and 'N' were swapped in seven cases. In 15 cases, digits were transposed (for example '0' as the first digit rather than the sixth) or apparently transcribed incorrectly (for example '8' rather than '5'). In five cases, the six digits of the survey NINO were '999999', suggesting a 'don't know' entry by the interviewer. In four of these cases, the NINO was reportedly derived from a payslip or other document and, in the other case, it was remembered with confidence. Indeed, in only two of the 32 cases was the respondent uncertain about the NINO. These examples suggest that the source of NINO measurement error is with the interviewer rather than with the respondent.

The lowest false-negative rates among the independent matching criteria were for matches by sex, date of birth and postcode: between 23.9% and 27.3%. The rate for matches by sex, date of birth, forename and family name was 30.6%, which is virtually the same as the rate for NINO matches (30.9%). The rate for NINO matches reflects the fact that a significant number of respondents did not supply a NINO—the problem of item non-response cited earlier. If all 62 of these cases had supplied a NINO and a genuine match had been made using this, then the NINO false-negative rate would fall substantially, to 19.2%.

The false-negative rate for criterion 3 would have been lower if there had been fewer mismatches on forename and surname. To illustrate the scope of pre-processing of name data for reducing this type of mismatch, consider the respondents with match pattern '11000'. Of the 68 cases, 39 non-matches by criterion 3 (and 4 and 5) arose because of differences in forename alone, and half of these appeared to be where the survey recorded a nickname. In seven cases, the forename was spelled differently, often only by one letter (for example 'Anne' *versus* 'Ann'). However, 16 non-matches arose because of differences in family name alone (typically not a simple difference in spelling) and 13 for other reasons, together comprising 43% of the 68 cases. Pre-processing therefore has some potential for improving match accuracy, but this potential is constrained (For an overview of US Census Bureau software for this and related tasks,

see Winkler, 2001). An alternative, or supplement, to pre-processing would be to relax the exact match on name using lookup tables based on common abbreviations or variants (e.g. surname plus initials), or other string comparison algorithms.

Choice of the best independent match criterion on the basis of linkage accuracy is clear cut according to Table 3. Criterion 2—matching by sex, date of birth and postcode—has both the lowest false-positive rate *and* the lowest false-negative rate. (It also had the highest raw linkage rate.) Observe that a shift to using hierarchical matching criteria reduced the false-positive rate associated with any match criteria involving the NINO (though the change is small). But false-positive rates did not fall universally. By contrast, false-negative rates for hierarchical matches were clearly smaller than for the independent matches, reflecting a decrease in the number of true non-matches (i.e. a fall in the numerator of the fraction). When matches from the five independent criteria were pooled, there were still some possible false-positive cases after our clerical inspections (cases with different addresses). The false-negative rate for pooled matching was zero (by assumption).

Summary and Conclusions

The positive conclusion of our study is that record linkage between household survey responses and administrative data is feasible, and even relatively simple and cheap matching procedures (as in our study) can yield good results when judged in terms of numbers of matches and their accuracy.

We have also provided new evidence about the choice of matching variables when linking respondents to household surveys with records from administrative databases. We have emphasized that the benefits gained from using unique personal identifiers like the NINO need to be assessed in the light of potential problems such as survey non-response to NINO requests (leading to higher false-negative linkage rates) and measurement error (leading to higher false-positive rates). Other personal variables common to the survey and the administrative data may also be used to create linkages, but they too have potential disadvantages. Not only is there potential measurement error, but also some information may differ in the two sources for legitimate reasons (e.g. names and addresses may refer to different dates in the two sources). Whether NINO-based matching, or matching by some other criteria, leads to higher and more accurate linkage rates is therefore a moot point.

Our study of linkages between ISMIE survey data and DWP benefit and tax credit records using five independent match criteria has highlighted these issues. The results suggest that linkages based on sex, date of birth, plus either postcode or first name and family name, yield a raw linkage rate as high as that for NINO-based linkages, and the linkages are relatively accurate.

Our hierarchical matching calculations underline the potential rewards to using additional variables for data linkage as a supplement to, or perhaps even instead of, NINO-based matching. For example, seeking a match on sex, birth date and postcode plus either NINO or forename and family name led to a raw linkage rate nearly as high as the pooled linkage rate derived when the results of all the independent matching

procedures were pooled. The fact that high linkage rates can be achieved without using NINOs is useful information for future linkage designers, given the additional burdens involved with collecting NINOs.

One route to improving linkage success rates is to raise the proportion of respondents who are willing and able to supply a NINO, and then to match using NINOs. However, since almost 90% of ISMIE respondents who gave their consent to DWP data linkage (a prerequisite for asking the NINO supply question) actually supplied a NINO, the potential for raising the NINO supply rate further is limited. To reduce false-positive linkage rates, NINO measurement error needs to be reduced. In our study most of the errors appear to have arisen from interviewer transcription error rather than respondent error. Since the potential for more sophisticated checking routines in CAPI scripts is limited, self-entry by a respondent might be a way to reduce this source of error.

How else might linkages between survey responses and administrative records be improved? Pre-processing of name and address data can help reduce inconsistencies between variables in household surveys and administrative record data. Our study underlined the potential of this for name data, but also suggested that its scope is constrained: a significant minority of non-matches (e.g. in surname) arose in ways that would not have been caught easily by cleaning algorithms. Our linkage rate for matches using address information would have been higher if the QAS^{TM} program could have been applied to the survey data as well as to the DWP data. However, since addresses in the two sources may refer to different dates for legitimate reasons, the application of software algorithms may have only a limited effect. The more that benefit file scan dates can be coordinated with the timing of the household survey fieldwork, the less that this will be a problem. Observe too that some of the problems described in this paragraph could also be mitigated if survey and administrative sources each contained histories of respondents' names and addresses, rather than a single observation for each.

It may be useful to investigate the relative merits of matching variables other than those used here. For example, the DWP Primary Data also includes telephone numbers for respondents. These numbers may also be routinely collected by survey agencies. There are, of course, potential problems as well: a significant minority of respondents may not have telephones, or change numbers relatively often (for example when changing mobile phone provider), and they may be subject to measurement error in the same way that NINOs are.

To get better data linkage results requires investment in matching technologies, not only in pre-processing software but also in the development of appropriate probabilistic matching algorithms. The returns to these investments will be greatest if the investments are coordinated between the major household surveys in order to take advantage of generic similarities in information collected that could also be used for matching.

Acknowledgements

This article derives from a project on 'Improving survey measurement of income and employment' (ISMIE), funded by the ESRC Research Methods Programme

(H333250031). We also benefited from ISER's core funding from the ESRC and the University of Essex. We are grateful to our ISER colleagues, especially Nick Buck, Jon Burton, John Fildes, Heather Laurie, Mike Merrett and Fran Williams, for their assistance in producing the ISMIE dataset. Helpful comments on earlier versions were provided by an anonymous referee, James Banks, Lucinda Platt and participants at an ESRC Research Methods Programme workshop on Data Linkage, 27 September 2004, London. We are also indebted to the Information and Analysis Directorate, DWP Information Centre, especially Catherine Bundy, Katie Dodd and Judith Ridley, for implementing the data linkages. The opinions expressed in this article are the views of the authors alone.

References

Böheim, R., & Taylor, M. P. (2000). *From the dark end of the street to the bright side of the road? Investigating the returns to residential mobility in Britain* (ISER Working Paper 2000–38). Colchester: University of Essex. Retrieved December 28, 2005, from http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2000-38.pdf

Calderwood, L., & Lessof, C. (2006, July 12–14). *Enhancing longitudinal surveys by linking administrative data.* Paper presented at the MOLS2006 Conference, University of Essex. Retrieved February 22, 2007, from http://www.iser.essex.ac.uk/ulsc/mols2006/programme/papers.php

Department for Work and Pensions. (2003). *Income-related benefits. Estimates of take-up in 2000/2001.* London: Department for Work and Pensions. Retrieved December 28, 2005, from http://www.dwp.gov.uk/asd/income_analysis/tu0001.pdf

Dex, S., & Joshi, H. (2004) *Millennium cohort study first survey: A users' guide to initial findings.* London: Centre for Longitudinal Studies, Institute of Education.

Gill, L. (2001). *Methods for automatic record matching and linking and their use in National Statistics.* (National Statistics Methodological Series No. 25). London: Office for National Statistics. Retrieved December 28, 2005, from http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=9224

Jabine, T. B., & Scheuren, F. J. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics, 2*(3), 255–277.

Jäckle, A., Jenkins, S. P., Lynn, P., & Sala, E. (2004). *Validation of survey data on income and employment: The ISMIE experience.* (ISER Working Paper No. 2004–14). Colchester: University of Essex. Retrieved December 28, 2005, from http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-14.pdf

Jenkins, S. P., Cappellari, L., Lynn, P., Jäckle, A., & Sala, E. (2004). Patterns of consent: Evidence from a general household survey. *Journal of the Royal Statistical Society: Series A, 169*(4), 701–722.

Jones, P., & Elias, P. (2006). *Administrative data as a research resource: A selected audit.* A report to the ESRC Research Resources Board. Retrieved February 22, 2007, from http://www.rss.org.uk/pdf/Admin%20Data%20selected%20audit%20report%20v2.pdf

Lessof, C. (2006, July 12–14). *Ethical issues in longitudinal surveys.* Paper presented at the MOLS2006 Conference, University of Essex. Retrieved February 22, 2007, from http://www.iser.essex.ac.uk/ulsc/mols2006/programme/papers.php

Lynn, P., & Lievesley, D. (1991). *Drawing general population samples in Great Britain.* London: Social and Community Planning Research.

Marmot, M., Banks, J., Blundell, R., Lessof, C., & Nazroo, J. (2003). *Health, wealth and lifestyles of the older population in England: The 2002 English Longitudinal Study of Ageing.* London: Institute for Fiscal Studies.

Noble, M., & Daly, M. (1996). The reach of disability benefits: An examination of the disability living allowance. *Journal of Social Welfare and Family Law, 18*(1), 37–51.

Olson, J. A. (1999). Linkages with data from social security administrative records in the health and retirement study. *Social Security Bulletin, 62*(2), 73–85. Retrieved December 28, 2005, from http://www.ssa.gov/policy/docs/ssb/v62n2/v62n2p73.pdf

Office for National Statistics (ONS). (2004). *Protocol on data access and confidentiality.* (National Statistics Code of Practice). Retrieved February 22, 2007, from http://www.statistics.gov.uk/about_ns/cop/downloads/prot_data_access_confidentiality.pdf

Plewis I., Smith G., Wright G., & Cullis, A. (2001). *Linking child poverty and child outcomes: Exploring data and research strategies.* (Research Working Paper No. 1). London: Department for Work and Pensions. Retrieved December 28, 2005, from http://www.dwp.gov.uk/asd/asd5/WP1.pdf

Winkler, W. E. (2001). Record linkage software and methods for merging administrative lists. (Statistical Research Report Series No. RR2001/03). Washington, DC: Bureau of the Census, Statistical Research Division. Retrieved December 28, 2005, from http://www.census.gov/srd/papers/pdf/rr2001-03.pdf