Understanding Society

Working Paper Series

No. 2020 – 02

January 2020

Integrated Data: Research Potential and Data Quality

Michaela Benzeval[1], Christopher R. Bollinger[2], Jonathan Burton[1], Mick P. Couper[3], Thomas F. Crossley[4] and Annette Jäckle[1]

[1] University of Essex

[2] University of Kentucky

[3] University of Michigan

[4] European University Institute

Understanding Society

Economic and Social Research Council

Non-technical summary

Integrating longitudinal survey and administrative data provides wide-ranging opportunities for new research. However, researchers need to consider sources of possible error in both types of data and new errors that may be created through the integration process for their subsequent analyses. We classify uses of integrated data into two broad types:

- Validation – in validation studies the two sources of data are compared to assess the quality of one or both, sometimes with an aim of improving survey design or administrative data processing, or of improving estimation.
- Enhancement – is an integration of data where a novel data set is created in order to enhance the *measurement* of a concept, enhance *content* by adding concepts not available in one source to the other, or enhance *representation* by 'filling in' missing units of data in one source from the other.

We provide illustrative examples of these different kinds of integrated data studies from research on income and employment, education and health.

Data quality is often considered using the Total Survey Error framework. We have adapted this to cover all three data processes: survey data collection, administrative data processing and linking the two data sources. There are two broad categories of errors that must be considered:

- Errors of representation, which occur if the sample of cases is not representative of the population of interest. This might be due to coverage errors, sampling errors, nonresponse errors, or adjustment errors.
- Measurement errors occur if the variables in the data do not correspond to the concepts of interest which might be due to specification errors, reporting errors, or processing errors.

We set out the ways in which each of these types of errors might occur in survey data, administrative data and through the linkage process.

Finally, we identify three priorities, which are currently the most under-researched and/or offer the most potential for methodological research on integrated data: 1) Errors of representation created in the linkage process, 2) best practice for validation studies, and 3) specification errors arising in enhancement applications.

Integrated Data: Research Potential and Data Quality

Michaela Benzeval (University of Essex)
Christopher R. Bollinger (University of Kentucky)
Jonathan Burton (University of Essex)
Mick P. Couper (University of Michigan)
Thomas F. Crossley (European University Institute)
Annette Jäckle (University of Essex)

Abstract: In this review we discuss how integrated survey and administrative data have been used for research on income and employment, education and health; where the future potential for integrated data lies; how we might think about the quality of integrated data; and what we currently know about different sources of error that can affect integrated data.

Keywords: data linkage, validation, survey data, administrative data, total survey error

Acknowledgements: We thank Ipek Mumcu and Dietrich Peña-Zuluaga for research assistance. *Understanding Society* is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. This project has been funded by the Economic and Social Research Council *UK Population Lab Innovation Development Grant* (ES/S016651/1). Christopher R. Bollinger was funded by a Leverhulme Visiting Professorship (VP1-2017-010).

Corresponding author: Thomas F. Crossley, European University Institute, tfcrossley@gmail.com.

1. Introduction

In response to the Great Depression and to changing economic and social science, the collection of data through surveys began to accelerate after the Second World War. The telephone and the computer made collection and processing of large surveys, drawn through formal probabilistic sampling techniques, tractable. By the early 1990s, myriad surveys were being collected to measure and track aspects of UK populations of individuals, households and businesses. These include the *Family Resource Survey*, the *Labour Force Survey*, the health surveys for England, Scotland and Wales, the *Annual Business Inquiry* (now *Annual Business Survey*), and the *General Household Survey*. Surveys such as these have the advantages that they cover a broad population, they measure key concepts in detail and they provide core information (for example, demographic information in individual and household surveys). These properties allow estimation of population characteristics so that researchers can uncover important relationships and trends in society.

Longitudinal surveys, such as the *British Household Panel Survey* and *Understanding Society*, provide unique opportunities to uncover important life cycle relationships, to identify long term consequences of social policies, and to observe dynamic relationships such as formation or dissolution of households or how parental characteristics impact children. A key development that has emerged over the last decade or two is the availability of administrative data for research. Computer technology has led to electronic record keeping in business, health, and government: the physical file folder was replaced by the computer file. These administrative and process data create a range of new exciting research opportunities, particularly if combined with survey data. For example, the Longitudinal Studies Review (Davis-Kean, 2017) identifies the benefits of such linkage as: 'improved representativeness in coverage, enhanced content and value of the data for research, and increased speed in creating insights' (p.6). The review also notes that the UK has both extensive administrative data and a world-leading set of longitudinal surveys, so that there is very significant potential value to combining survey data with administrative record data in the UK.

The types of administrative data we are considering here include vital statistics (birth and death records), health records, financial reports, educational transcripts and government records on benefits, earnings, and income. These records are obtained through processes such as birth or death certification, applying for a mortgage, applying for or receiving social benefits, filing tax returns, or seeing a physician for health care. They may include direct measurements (such as height and weight taken by a clinician) or responses of subjects themselves (such as filling in application forms). They offer the opportunity to study all of the population who participate in these processes.

Administrative, or process records, have a number of important drawbacks. First, their coverage of the population is often neither complete nor random, and is not designed for the purposes of research or policy evaluation. Rather their coverage of the population is related to the processes the records cover. For example, while data on those who receive social benefits such as Housing Benefit are collected, they do not contain any data on those who do not receive (or apply for) these benefits. This prevents investigation of the process of moving into or out of these programmes, and prevents comparisons of important outcomes between participants and similar non-participants.

Second, administrative records seldom contain the rich demographic and social information that social scientists require for research. While some contain age and gender, and perhaps ethnicity and postcode, which enables the addition of area data, few if any contain educational attainment (even education transcripts may not, as education is attained across different levels of organizations), family history, or parental achievement for example. Administrative records focus on the domain in which they are collected, for example tax records have information on earnings and health records on health diagnoses and service utilisation. There are no current administrative records that cover multiple domains, although there is the potential to combine different administrative records to provide a more comprehensive picture of people's lives. Moreover, administrative data do not contain more subjective – but often critical – types of data such as happiness or expectations. In contrast, social and health scientists have accumulated experience in collecting such constructs in surveys.

Third, administrative data by definition are collected as part of other processes so that data quality may be variable for a variety of reasons. Concepts may not be systematically collected or coded (for example across doctors or benefit offices) or may change over time as administrative requirements change. There may be incentives within the process which lead to misreporting of key concepts (for example under-reporting of taxable earnings to taxing authorities, or over-reporting of earnings to lending agencies). The operating definitions in process data may not match the social or economic concepts researchers are interested in measuring (billing codes in medical records may not match actual diagnoses, annual or quarterly earnings may be collected rather than the wages or pay period).

Fourth, access to individual-level administrative records for research purposes, especially when wanting to combine information, currently faces numerous barriers to effective access and use on a wide scale for research (Davis-Kean, 2017). We do not focus on these barriers here, but acknowledge that the examples of beneficial research below are still, and will continue to be, relatively rare until these institutional barriers can be overcome.

Survey data too have drawbacks. First, it is difficult for any survey to measure every possible concept or variable well. The longer the survey, the less likely respondents will participate and complete a full interview. Thus, we have a *Labour Force Survey* to measure work and a *Family Resources Survey* to measure income. Surveys that cover multiple domains (such as *Understanding Society*) are necessarily limited in their depth of content in each domain. Second, there is increasing concern about some aspects of the quality of survey data, notably selective response (see for example Meyer *et al*, 2018) or that samples were not designed to cover key policy-related population subgroups such as small areas, populations living in institutions, or the homeless.

Combining administrative and survey data offers the possibility of exploiting the strengths, and mitigating the weaknesses, of both. In this review, we focus particularly on combining survey responses for basic units (e.g. from people, households, or firms) with administrative records on the same or similar units. Data combination might involve linkage, where subsequent research is based on the intersection of the units responding to the survey and the units captured in administrative records. For example, income data from tax records might be combined with labour market details from survey data to investigate productivity. Alternatively, it might involve augmentation, where subsequent estimates are based on the union of these sets of units (to improve coverage of a target population, for example). For example, missing survey income data on high earners may be imputed using data from tax records. In either case we refer to combined data as *integrated data* and we give examples of several types of integration below.

In this review we discuss:

- How integrated survey and administrative data have been used to date, and where the future potential for integrated data lies.

- How we should think about the quality of integrated data, and what we currently know about the quality of integrated data.

An important research agenda is to consider how *integrated data* are best used to improve estimation of important relationships and models in the social and health sciences. Both survey and administrative data are subject to measurement error and missing data, but the processes generating these error sources may be very different. Moreover, while the process of combining or linking datasets may help overcome some data problems, it can in turn generate new problems. Linkages can be probabilistic (based on common information such as name, birthdate, address) or deterministic (based on government IDs such as National Insurance or NHS numbers). Data quality issues for integrated data can arise due to poor matches or lack of records for some types of individual (for example the unbanked) or administrative records which no longer correspond to individuals in the target population (such as financial records of those who have deceased or emigrated).

In the next section we discuss two broad ways in which integrated data have been used to date. The first is often referred to as *validation* data, and we will adopt that term. In this category, survey and administrative data are used together, but one source is typically used to draw inferences about the quality of the other. These quality inferences are then used to assess the reliability of scientific inferences drawn from the first data source, to correct estimates made with one data source, or to guide improvements in data collection or processing.

In the second broad category, which we term *data enhancement,* data from administrative and survey sources are used to "complete" each other. This category includes cases where the survey and administrative data contain information on different concepts so that the combined data have wider content for research purposes. It also includes cases where a measure in one source is replaced by a superior measure of the same construct (for the same units) from the other source. For example, administrative data may provide a measure of a variable that is difficult or impossible to measure in a survey (or a survey may provide a measure of a concept that is not recorded in administrative data). Finally, the survey and administrative data may contain different units so that the combined data have superior representation of the population of interest. The survey may capture types of units that are missing from the administrative data (such as program nonparticipants) or, of course, administrative data may capture types of units that are missing from surveys (such as very high-income individuals). In all cases this data enhancement results in an essentially new (enhanced) research data set, which is then used to estimate quantities, relationships or models of interest.

For both *validation* and *enhancement,* we provide illustrative examples, drawn particularly from studies of labour markets and income, health, and education.

Box A: Key Terminology

Integrated data – any combined use of survey data with administrative or process records for the same or similar units. For example, survey responses from individuals with administrative records on the same or other individuals.

Data Validation – Survey and administrative data are linked or appended. However, the focus is on using one source to assess the quality of the other, or to use comparisons between the two to improve data collection or processing, or estimation.

Data Enhancement – survey and administrative data are integrated so that effectively a new research data set is created. The survey and administrative data contain information on different concepts, or different units, or a measure in one data set is replaced by a superior measure of the same construct from the other.

Total Survey Error (TSE) and Total Error (TE) – frameworks for thinking about the sources and effects of different kinds of error in surveys (TSE) or in any type of data used for the estimation of quantities, parameters or relationships, including administrative or integrated data (TE).

In the third section of this review we take up the question of the quality of integrated data. We discuss how the concept of Total Survey Error (TSE) can be expanded to “Total Error” (TE), and how the latter can be applied to integrated data in particular. Total Survey Error focuses on the sources of errors in surveys, and their impact on estimates. The errors include errors in coverage of the population, sampling error, non-response error, specification error, and measurement error. The Total Error concept applies the same idea to any data used for estimation. In the context of integrated data, this will include the elements of TSE applied to the survey, elements of TSE such as measurement error and coverage applied to the administrative data, and new sources of error generated by the process of combination itself. This provides a framework for thinking about the quality of integrated data and when and how survey and administrative data can best be combined for estimating quantities and relationships of interest.

2. How has Integrated Data been used?

Integrated data takes many forms. The broadest possible definition could include cases where individual-level data are merged with contextual information from an administrative source at a geographic or other aggregate level. The Longitudinal Studies Review highlights the value of such combinations of data enabling researchers to address ‘innovative research questions, particularly on separating factors on the micro-level of individuals, families and households from environmental factors on the meso-level of neighbourhoods, schools and work organisations, and the macro-level of opportunity structures and institutions’ (Davis-Kean, 2017, p.25). For example, merging labour market participation data to information on the region in which the survey respondent resides; merging survey information on hospital use to the characteristics of the nearest hospital for a respondent; or merging school quality data to information on students and their families in a survey. However, in this review we focus on a narrower definition where information is brought together from survey and administrative records at the same unit level. For example, individual surveys responses might be combined with earnings data from tax records, clinical data from NHS health records, or transcript data from schools, or firm-level survey responses might be combined with firm-level VAT receipts. We choose this narrower definition because the broader cases – linking contextual, often geographic, administrative data to survey data – have long been available and utilized by researchers in many contexts.

As described in the introduction we distinguish two broad categories of integrated data, *validation* and *enhancement*. In the former differences between the sources are employed to assess the reliability of results or to inform and improve data collection or estimation methods. In enhancement, information is combined from multiple sources to generate effectively new research data, which are used to estimate quantities, relationships and models of interest. Within these categories we can further distinguish various subcategories by the specific research goal. We summarize these in Table 1.

Table 1: Uses of Integrated Data

Type of Study	Sub Category	Goal
Validation		Comparison of survey and administrative data.
	Descriptive	Simple description of differences, often to highlight reliability of results.
	Data Improvement	Using differences between sources to improve the collection or processing of either survey or administrative data.
	Estimation Improvement	Using differences between sources to generating better estimates of quantities of interest.
Enhancement		Integration of the data sources to effectively create a new research data set.
	Enhancements of Measurement	Replacing variables in one source with a superior measure (either from the other source, or some combination). Measurement is improved, but content is not widened.
	Enhancements of Content	Adding variables from one source that were not collected in the other. Content is widened.
	Enhancements of Representation	Filling in coverage or response gaps (missing cases) in one source with records from the other.

Validation Studies

The primary goal of validation studies is to use one data source to evaluate the other. Most often the administrative data are considered to be the standard against which the survey data are validated. However, this is not always the case. For example, Browning and Leth-Petersen (2003) use budget survey data to validate a measure of household consumption derived from tax records on income and changes in wealth.

We further distinguish three subcategories of validation study, differentiated by their goals. The first, and most basic, subcategory reports simple descriptive studies; the goal is simply to provide a descriptive evaluation of the data quality as it exists, often to assess the reliability of results based on one of the survey or the administrative data. The second group of validation studies are those that aim to use the comparison of the sources to improve data collection. Most often comparison with administrative data is used to improve survey data collection, but comparisons with survey data can feedback to the collection or processing of administrative data. Finally, the third uses information from multiple data sources in the estimation of the quantities and relationships of interests, but in a way that stops short of full integration of the data.

Descriptive

This is perhaps the most common form of validation study. Some of the earliest examples of this kind of study were done in the U.S., with a validation sample specially drawn for the *Panel Survey of Income Dynamics* (PSID). A sample of respondents to the PSID were matched to their employment records at a large manufacturing firm, as described in Morgan (1989). Using these data, Mathiowetz and Duncan (1988) examined how reports of unemployment were recalled by participants. They found that survey respondents often ‘telescope’ or shorten the length of the spell when recalling past unemployment. Using the same data, Rogers *et al* (1993) examined how reports of earnings and hours worked differed. They found that while there was some regression to the mean for annual earnings, these reports were typically better than reports of hourly or pay period earnings. Reports of hours worked were particularly problematic, and led to significant bias in estimates of hourly earnings when constructed from annual earnings and hours worked.

Moore and Marquis (1989) and Marquis and Moore (1990) matched a subsample of the *Survey of Income and Program Participation* (SIPP) to administrative records from three

U.S. states to examine nine programs (Aid to Families with Dependent Children (AFDC), Food Stamps, Unemployment Compensation, Workers Compensation, Civil Service Retirement, Pell Grants, Social Security, Supplemental Security, and Veterans Compensation). They find that errors in reporting participation varied widely from program to program. They find that for some programs, such as Social Security, receipt and amounts were relatively accurate, while for programs such as AFDC and Food Stamps, significant underreporting of participation was present. Card *et al* (2004) examine the Medicaid coverage in California responses to the SIPP finding approximately 10% underestimation of program use. Cohen and Carlson (1994) examine reports of expenditures for medical procedures in the *National Medical Expenditure Survey* compared to a subsample of linked provider records. They find substantial agreement on expenditures for those who report the existence of the expenditure. They also find that the larger the percentage of the expenditure that is out-of-pocket (paid directly by the respondent) the greater the likelihood of reporting the expenditure, and the greater the accuracy of the report.

In more recent work, rather than focusing on a special subsample, existing surveys have been matched completely with the full universe of administrative data. Bound and Krueger (1992) and Bollinger (1998) both use the 1979 match of the *Current Population Survey* (CPS) to records of the U.S. Social Security Administration. They both find significant differences between survey reports and administrative records; in particular, they find support for a negative relationship between measurement error and the level of earnings. Roemer (2002) examines both the SIPP and the CPS compared to tax records and finds as well that for low earners earnings reported in the surveys are higher than earnings recorded in the administrative record. They attribute this to underground earnings. Abowd and Stinson (2013) examine earnings in the SIPP and allow for measurement error in both the survey and the Social Security Administration records. They find both sources had measurement error and estimated that both had similar amounts of measurement error. Bollinger *et al* (2018) similarly examine measurement error in earnings reports in the CPS and draw similar conclusions that measurement error in the administrative records may bias estimates of measurement error in the survey data.

Meyer *et al* (2018) also find substantial underreporting of participation in the Supplemental Nutrition Assistance Program (SNAP) in the CPS. Overall, they find that SNAP participation is under-reported by as much as 23% and that misreporting differs by race and family size. Authors such as Mellow and Sider (1983), Rogers and Herzog (1987), and Bound *et al*

(1994) examine measurement error in earnings and employment. Results in Bound *et al* (1994) for example, suggest that the errors in reports of earnings are negatively correlated with true earnings, and may bias estimation of returns to education and experience downward. Battistin *et al* (2014) examine how misreports of education qualifications impact the estimates of returns to education. They find that both transcript data and self-reported data contain errors, and that on net, the estimates obtained of returns to qualifications over-estimate the return.

Pascale, Roemer, and Resnick (2009) find 36.2% of true Medicaid enrolees fail to report enrolment in the CPS, but there were very few false-positives (non-enrolees erroneously reporting enrolment). They find that the results are likely due to the long recall period (12 months).

In the health field a range of studies have compared prevalence and associations in survey data combined with administrative data to assess misreporting, misclassification and missingness. For example, Griffiths *et al* (2017) linked the *Millennium Cohort Study* to General Practitioner (GP) records in Wales to compare levels of asthma as reported by parents and recorded in GP records. Parents appeared to over-report compared to administrative records, although differences reduced as children aged and if a broader definition of respiratory disease codes in the GP records was considered. In a Dutch study comparing data from the Perinatal Registry with questionnaire data (for adult survivors of childhood cancer), information on birthweight was highly correlated between the two sources. However, not all pregnancies reported could be linked to the Registry, the completeness of information in the Registry could not be checked as there was no reporting of missing records, and for some outcomes there was a higher level of missingness in the Registry than self-report data (Overbeek *et al*, 2013). In a study of the association between the duration of breastfeeding and teenage IQ in the *Avon Longitudinal Study of Parents and Children* (ALSPAC), Cornish *et al* (2015) use administrative data on education attainment from the National Pupil Database (NPD) to investigate what kind of patterning of missingness the IQ score has and hence if the survey analyses are biased. Despite including a wide range of covariates from the survey data, the inclusion of the linked data showed that there were differences between participants with and without IQ data, and hence analyses relying solely on survey data would be biased.

Improving Data Collection and Processing

The second type of validation study seeks to improve on data collection. One such project was the Income Survey Development Program (ISDP) (see Vaughan *et al*, 1983 for a broad discussion) which led, in part, to the design of the SIPP. Work on Aid to Families with Dependent Children (AFDC) (Klein and Vaughan, 1980; Goudreau *et al*, 1984), supplemental security income (Vaughan, 1978), social security (Vaughan and Yuskavage, 1976) and veterans' payments (Vaughan *et al*, 1983) are investigated more specifically. Overall, this program drew small samples of individuals from various transfer programs in the United States, and used a standard survey instrument to assess the quality of data obtained in face-to-face interviews. Findings were often quite general across the programs. Perhaps the most important was that shorter recall windows helped in reducing recall error. This led, in part, to the design of the SIPP having four-month reference periods (interviews are done three times per year). Other findings include that explaining problems in recall, such as failing to report very short spells from the beginning of the recall period, can alleviate these problems to some degree. Many of these approaches have been widely included in surveys since.

Kreuter *et al* (2010) link the first wave of the German Panel Study "Labor Market and Social Security" to administrative data on welfare benefits, employment status, age and citizenship. They focus on differences in recruitment effort to examine both response rates and response error. They find that while additional recruitment effort reduced non-response error, there was some increase in measurement error, so that additional recruitment efforts led to an overall increase in total survey error.

Perhaps the most ambitious validation studies of this type use administrative data in conjunction with experiments in survey design and implementation. For example, Lynn *et al* (2012) and Jäckle and Eckman (2019) conducted survey experiments with different methods for reducing measurement error in longitudinal data and used linked government administrative data to test which method produced the most accurate data. Kreuter *et al* (2008) examined the influence of the mode of data collection on both non-response bias and accuracy of reporting on sensitive topics, by combining a mixed mode experiment with linked administrative data. The goal in these studies is to use the administrative data as a "gold standard" in order to choose between alternative survey procedures.

Improved estimation with validation data

Finally, validation data is sometimes used to improve estimation. One approach is to use the validation data to provide estimates for key parameters (such as error rates or error variances in the primary source) which can then be employed by analysts to adjust estimates. This typically involves assuming that validation data (most often the administrative data) is a "gold standard." For example, Bollinger and David (1997) estimate the rates of misreporting of Food Stamp program participation using a subsample of the SIPP matched to administrative records of participation. They find a strong relationship between income and failure to report participation. They use the results to then adjust estimation of food stamp program participation, finding that failure to adjust for misclassification leads to underestimates of the impact of asset holdings on participation.

While work of this type often assumes that one source (usually the administrative data) is a "gold standard", improved estimation does not require this. It is possible to improve estimates with validation data even if both source data sets contain measurement error. For example, Cajner *et al* (2019) use employment data from a survey (Current Employment Statistics from the U.S. Bureau of Labour Statistics) and microdata from a payroll processing firm to estimate U.S. employment growth. They argue that both sources contain measurement error, but that by using a statistical model that derives optimal predictions from the two sources, they obtain estimates that are superior to what could be obtained from either source alone.

This type of validation study can address coverage problems, as well as measurement problems. A nice UK example is the work of Jenkins (2017). Jenkins uses survey data (from the Family Resources Survey) and administrative tax records to study changes over time in income inequality. He argues convincingly that household surveys provide poor coverage of very-high-income households and individuals, so that the samples are effectively truncated (on the right tail). He addresses this by combining inequality estimates for the bulk of the distribution from the survey with inequality estimates for the top tail of the income distribution from the tax data. Note that the survey data are preferred for low- and middle-income individuals, partly because many lower income individuals are not required to file tax returns in the UK. Using both survey and tax data in the estimation process substantially alters conclusions about the time path of income inequality in the UK. Jenkins estimates that between 1996/7 and 2007/8 the Gini coefficient for individual gross income rose by 7-8%. Estimates based on survey data alone suggest a fall of 5% for the same period.

Data Enhancement

The second broad category of application of integrated data is when administrative data are combined with survey data to create new data that is superior to either of the individual sources. We further differentiate between three types of enhancement: *enhancements of measurement*, *enhancements of content* and *enhancements of representation*. Enhancements of measurement and content are those cases where researchers are interested in the set of units captured by a source data set (administrative or survey) but want to add a concept or improve a measurement. This will involve data linkage.

Enhancements of representation are those cases where administrative and survey data are combined to address errors of coverage or selective nonresponse, or more generally to improve representativeness. Here integration may fill in "missing units" in one data set with the other. The integrated data to be analysed will often be the union of the units captured by the survey and administrative sources, whereas with enhancements of content or measurement, analysis is naturally restricted to the (linked) intersection of the two sets of units.

Enhancement of Measurement

We first consider enhancements of measurement. Data combination can replace a measure of a concept in the first data set with a superior measure from the second. A common example of this type is the addition of administrative income data to surveys. The administrative income data is superior to what might be collected in terms of accuracy, granularity, completeness, or historical coverage. For example, work by Juhn and McCue (2016) links survey data from the SIPP with earnings data from Social Security records. In the integrated data the SIPP provides detailed marital and birth histories for the women, while the matched Social Security data provides long earnings histories that would be difficult to capture accurately and completely in a survey. Juhn and McCue study how marriage and childbirth have impacted the earnings of women compared to single women. They find that over cohorts the marriage earnings gap (married women earn less than their unmarried counterparts) has diminished, but the birth earnings gap has remained constant.

In the UK, survey data from the *English Longitudinal Study of Aging* (ELSA) linked to National Insurance data have allowed researchers to examine how past earnings impacted individuals during their retirement (Bozio *et al*, 2010; Bozio *et al*, 2013; Bozio *et al*, 2017;

Crawford *et al*, 2014 and Crawford and O'Dea 2014). For example, Bozio *et al* 2017 uses a measure of a couple's lifetime earnings derived from survey data where both partners are in the study, and national insurance and survey data where one of the partners is missing, to estimate the association between couples' lifetime earnings and wealth accumulation.

Enhancements of Content

Enhancement can alternatively involve the addition of measures of entirely distinct concepts that are not measured at all in the source data. Probably the most long-standing and common example of this is adding data from administrative mortality records to survey information. This approach has been used extensively in health research (e.g. Doll and Hill, 1954; Haan *et al*, 1987; Marmot and Shipley, 1996; Burström and Fredlund, 2001; Whitely *et al*, 2014). For example, Doll and Hill (1954) surveyed GPs about their smoking habits, and then followed up by linking to National Mortality Registry data, providing the first evidence of the detrimental effects of smoking on health. Similarly, there is a significant literature linking survey data with administrative hospital records. For example, in Scotland where administrative health data has been used successfully for research for some time, a number of older cohort studies have been very successfully followed and analysed for decades by passively following up on earlier life surveys through record linkage (e.g., the Midspan Studies – see Hart *et al*, 2005). More specifically, in other integrated analyses, researchers have used the richness of data on participants' characteristics and health from survey data combined with hospital records in a wide range of applications. Examples include: the impact of childhood obesity on hospital admissions (Griffiths *et al*, 2019), preventable hospital admissions (Falster *et al*, 2019), and hospital costs (e.g., Geue *et al*, 2015).

The series of papers by Stinebrickner and Stinebrickner are a useful example of this in the education field, and also serve as a reminder that it can be the administrative data, rather than the survey data, that is augmented. These authors were concerned with understanding college dropout (university drop out in UK terminology) among low income students, which is an important problem, particularly in the United States. In their first paper (Stinebrickner and Stinebrickner, 2003) they had administrative data in the form of student records from Berea College (a liberal arts college in the U.S. which primarily serves students from lower income backgrounds). While this provided novel associations of student characteristics with subsequent dropout, the mechanisms driving dropout (and hence generating these associations) remained largely a black box. This was because key variables in plausible models of the dropout decision, such as the credit constraints faced by students, or their expectations of academic success and of earnings, were not captured in the administrative data. They then designed a survey using the administrative data as a frame to generate an integrated data set. The survey enhanced the administrative data by providing measures of key variables such as expectations and credit constraints. This allowed subsequent research which provided a much-enhanced understanding of the dropout decision (see Stinebrickner and Stinebrickner, 2008 and 2012).

In the UK, the Department for Education calculates value-added measures of school performance using administrative data. A possible concern with these calculations is that the administrative data do not contain measures of background variables that are known to be important determinants of educations outcomes. If such variables vary across schools, then the value-added calculations could suffer from important omitted variable bias. Dearden *et al* (2011) use linked survey and administrative data to assess such biases, particularly focussing on mothers' education as a potential omitted variable. They exploit a linkage that exists for a particular cohort of students to add mothers' education from the *Longitudinal Survey of Young People in England* (LSYPE) to the NPD (the administrative data source). Value-added calculations on these enhanced data indicated that failure to control for mothers' education can indeed lead to significant biases in value-added comparisons across schools.

Other excellent examples of enhancement of content in the education field include Baron and Cobb-Clark (2010), Broecke (2012), Rouse *et al* (2013), Dale and Kreuger (2014), and Cobb-Clark *et al* (2015).

Enhancement of Representation

Integrated data can also be used to enhance coverage. We discussed above a validation study (Jenkins, 2017) that addressed the poor coverage of high-income households and individuals in household surveys by combining *estimates* for inequality for different parts of the income distribution from the survey data (for the bulk of the distribution) and from tax data (for the top tail). Burkhauser *et al* (2017) address the same problem; however, in contrast to Jenkins, they combine *data*, rather than *estimates* (and then use the combined data as the basis for estimation). Thus, in our typology it is an enhancement study (in particular, an enhancement of representation). Their main approach is the same as the "SPI adjustment" that DWP makes to the *Family Resource Survey* data before using the resulting enhanced data to produce the official income statistics in the UK ("Households Below Average Income"; see Department for Work and Pensions, 2015). Like Jenkins (2017) this approach uncovers rising income inequality that is missed by analyses of survey data alone.

In another example Gray *et al* (2019) employ *Scottish Health Survey* data alongside administrative hospital records to improve estimates of drinking prevalence derived from the survey, which may be biased by the omission of survey non-responders. Gray *et al* (2019) use the administrative data to estimate the differences in demographics and hospitalisation between respondents and non-respondents, create synthetic observations for non-respondents and then impute drinking prevalence to these synthetic non-respondents. Thus, they improve coverage by using administrative data to add survey non-responders back into the analyses.

Finally, Sakshaug and Antoni (2019) take a very different approach to using integrated administrative and survey data to address limitations which arise due to errors of survey coverage and representativeness. Rather than using administrative data to replace or impute missing units, Sakshaug and Antoni (2019) explore how linked administrative data can be used to improve weighting adjustments for nonresponse in a survey. Thus, in this case the enhanced data do not have additional units, but rather new weights. They find that weighting adjustments incorporating variables from the administrative data are superior to adjustments based on survey paradata alone.

3. The Quality of Integrated Data

In the previous section we saw that integrated data are often used to address errors or other limitations of a source data set. This may be through improving collection, processing or estimation in a validation study, or through the creation of a new enhanced data set from the combination of survey and administrative sources. These data are, by design, superior to the source data sets, at least in some dimensions. However, particularly when new data are created through enhancement, it is important to remember that integrated data will also be subject to errors. First, any errors in the source data sets that are not addressed by enhancement will be inherited by the integrated data. For example, if we link administrative data to survey records to provide a superior measure of income, the integrated data will inherit any coverage errors in survey data. Equally, administrative data will also have errors of various types, and these too can be inherited. In addition, new errors may be introduced by the process of integration. It is therefore important that researchers think systematically about the sources of error in integrated data and their consequences for estimation.

Harron *et al* (2017) and Sakshaug and Antoni (2017) contain excellent discussions of the types of errors that can arise through the process of data linkage. Harron *et al* propose a number of ways in which the degree and importance of linkage error can be assessed. Gilbert *et al* (2018) provide guidelines for the types of information that data linkers should provide to researchers. However, what we argue here is that the quality issues around integrated data should be considered more broadly than just errors arising in linking additional data to set of units. First, not all integrated data involves linkage. As discussed in the previous section, integrated data can also involve the addition of units to address coverage or response problems. Second, as noted above, errors in integrated data may be inherited from the source data sets, as well as in arising from the processes of data integration. A similar point is made by Hand (2018) in the context of UK Small Area Income Statistics, which combine survey data with a number of administrative data sources (including claimant counts).

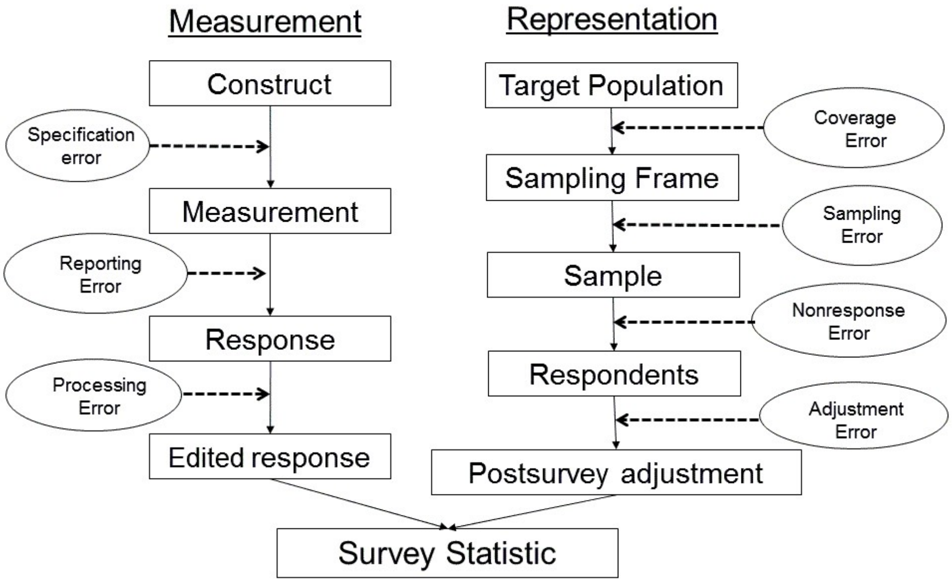
In this section we therefore lay out a framework for thinking about all of the potential sources of error in integrated data. Our framework builds on the well-known Total Survey Error framework for survey data, and on recent extensions of that framework to other data types. We begin by briefly reviewing the Total Survey Error Framework.

The Total Survey Error framework

Data quality is driven by the processes that generate the data, and by potential errors that can occur at the different stages of these processes. In survey methodology, the Total Survey Error (TSE) framework is widely used to describe the kinds of errors that can occur in surveys and their potential impact on statistical estimates (see e.g. Groves, 1989). Errors are viewed in a statistical sense as deviations from the desired outcome, rather than as mistakes. Errors can be of two broad types: biases introduce systematic errors into estimates, while variable errors affect the precision of estimates. Errors are also classified into two broad classes: errors of representation (selection bias) and measurement errors.

Figure 1 illustrates the survey process and associated error sources. Errors of representation occur if the sample of survey respondents is not representative of the population of interest. This might be due to coverage errors, sampling errors, non-response error, or adjustment errors. Measurement errors occur if the variables in the data do not correspond to the concepts of interest. This might be due to specification error, reporting error, or processing error.

Figure 1: The Survey Process and Error Sources That Can Occur at Each Stage



Source: Adapted from Figure 2.5 in Groves *et al* (2009)

Many of these kinds of errors apply equally to data generated for administrative purposes. Hence, while the focus of TSE is on survey errors, the framework has recently been expanded and adapted to accommodate errors in other data sources, such as so-called “big data” and administrative data (e.g. Biemer *et al*, 2017). As such the framework is now being termed the Total Error (TE) framework to reflect its broader applicability (see e.g. Lavrakas, 2013; Biemer, 2014). Connelly *et al* (2016) and Groen (2012) also refer to the TE framework in the context of administrative data.

An extremely important point is that administrative data, like survey data, have data generating processes, errors can enter through those processes, and indeed, steps in those processes often include human actions (Connelly *et al*, 2016; Hand, 2018). The UK Statistics Authority (2014) has noted: *‘we have been surprised by the general assumption made by many statistical producers that administrative data can be relied upon with little challenge, and, unlike survey-based data, are not subject to any uncertainties’* (UK Statistics Authority, 2014).

Adapting the TE Framework to Integrated Data

Our key argument is that errors may be inherited from both data sources, as well as arising through the process of integration, and that the TE framework is useful for understanding the errors in order to help mitigate their impact on inferences made from integrated data.

Table 2 summarizes the sources and types of errors that can arise in integrated data. The rows of Table 2 follow the TE framework in classifying errors into errors of representation and errors of measurement, and then further dividing those classes. The columns of Table 2 address, from left to right, errors entering through the survey data, through the admin data, and through the process of integration. Below, we discuss these in turn.

Table 2: Errors in Integrated Data from survey data, administrative data, and data linkage

	From Survey data	From Administrative data	Via Linkage
Errors of representation			
Sampling error	Increased variance due to sampling Bias if certain units of the population systematically excluded	Sampling sometimes necessary to handle large volume of data. Bias if non-probabilistic sampling or probabilistic with unknown sampling probabilities	Sampling properties unknown when data to be analysed are the intersection of units in the two sources
Coverage error	Mismatch of the sampling frame and the population of interest (e.g. exclusion of homeless or those living in institutions from address-based samples)	Population covered is defined by who is involved in the process that generates the data (e.g. unbanked not included in financial transactions data, those who use private health care in UK not included in administrative NHS hospital records).	The intersection of units captured by survey and administrative data may cover less of the population than either source individually
Non-response error	Bias if non-respondents differ from respondents in key characteristics	A unit that should be in the administrative data is not (e.g. an individual failing to file a required tax return; or failing to seek help for a medical condition).	Bias if those who do not consent to data linkage differ from those who do on key characteristics Bias if error in linkage variables leading to linkage failure is associated with characteristics of interest
Measurement error			
Specification error	Questions are designed to measure concepts of interest, but not all research questions will be foreseen at design stage	Possible mismatches in concepts of interest (e.g. total income or taxable income) or units of interest (e.g. individuals vs households)	Differences in units of aggregation or temporal units between the survey and admin data. Errors due to delays in updating admin records

Reporting error	Due to respondent errors (e.g. recall problems or deliberate), or influence of the interviewer, questionnaire, or mode of data collection on responses	Due to errors made by individuals completing forms (e.g. company administrators providing information about individuals; individuals providing information about their income in tax returns), systematic biases may be introduced by fee structures (e.g. in UK GP records)	Incorrect linkages may lead to incorrect information being assigned to units
Processing error	Coding or classification errors, transcription errors with paper questionnaires	Errors in transcribing data from paper forms; errors in data processing and reduction	Errors in data processing and reduction

Errors Entering Through the Survey Data

There are three key types of errors of representation in surveys. *Sampling error* arises from the need to select a sample rather than studying the full population. Sampling error usually increases the variance (reduces the precision) of estimates, but typically has little effect on bias. *Coverage error* arises from the fact that segments of the target population are systematically excluded from the sampling frame. Examples are the exclusion of non-Internet users in online surveys or the homeless and institutionalised populations in household surveys. *Non-response error* arises through that fact that not all sample members participate in the survey, either because they could not be contacted or because they decline to participate. To the extent that those excluded differ from those included in the survey, non-response error can introduce bias in estimates. If a survey sample is used as the basis for linking, sampling errors affect the combined data set. Similarly, unless the process of integration removes them (as in some of the enhancement studies discussed in the previous section) survey coverage and non-response errors will be inherited by the integrated data.

The second broad class of errors are related to the process of measurement. *Specification error* occurs if the concepts measured by the data or method do not coincide with the concepts of interest for research purposes. Given that surveys are “designed” data (i.e. they are designed to measure the concept of interest; see Groves, 2011), specification errors in surveys are minimised at the design stage. However, it is certainly true that no survey can foresee all possible applications and concepts required. In many cases, specification error arises ex-post when researchers have interests unforeseen by survey design. There is also a tension in survey design between interview length and covering all possible concepts. *Reporting or response errors* are typically the biggest source of measurement error in self-report surveys. Respondents may misreport (or fail to report) "correct" information for a variety of reasons, including problems understanding the question, recalling or retrieving information from memory or records, or altering the response (e.g. socially desirable responding, satisficing, etc.). Responses can also be influenced by the interviewer, the questionnaire, or the mode of data collection. Finally, *processing errors* can arise during the process of coding, classifying and preparing the data for analysis.

Errors Entering Through the Administrative Data

Starting with errors of representation*, sampling errors* are often viewed as not relevant to administrative data, as the data are often available for the full set of participants. Increasingly, however, because of the size of administrative data sets, sampling is being used. For example, the Credit Reference Agency data studied by the UK Financial Conduct Authority is a 10 percent sample of the full file. As with surveys, sampling error in administrative data reduces the precision of estimates, but is not normally a source of bias.

Other types of errors of representation can and do occur in administrative data and, as with errors of representation in a survey, can be an important source of selection bias in estimates. Errors of representation can occur because some units are not covered by design (for example, in the UK not all individuals are required to file tax returns, and in most countries not all households will receive pensions and benefits). Alternatively, units that should be captured by the data may not be. Further examples included the unbanked or under-banked (not included in credit records or other financial datasets), those with education from abroad (who may be absent from education administrative records data), and those with no interaction with the National Health Service (e.g. people who do not register with a GP, often those who are highly mobile or homeless; those who use private health care or those with symptoms who choose not to seek medical assistance, see Herrett *et al*, 2015). Administrative data seldom cover populations such as refugees or undocumented immigrants, while recent evidence (Bollinger *et al*,2019 suggests that surveys may cover this population. Depending on how the administrative data are compiled, i.e. which data sources they draw on, the data may or may not represent a segment of the population that is well defined. For example, credit reference data might be representative of credit products held in a country, but need not be representative of people currently living in the country.

Turning now to errors of measurement, *specification error* is potentially more likely in administrative data than survey data as administrative data are typically not constructed with analytic purpose in mind and are usually collected for a purpose other than statistical estimation. There are two potential sources of specification error. *Unit of analysis errors* occur when the administrative data are collected about different units than the analytic goals. Examples include data on persons versus households or benefit units, or account- or mortgage-holders (may be jointly held accounts).

Conceptual differences refer to concepts defined for administrative purposes that may not match the analytic goals. In U.S. medical records, billing codes may not match medical diagnoses or self-reported conditions. For example, in primary care in the UK although there is a standard coding system, GPs complete this information during consultations for medical rather than research purposes and often write free text notes for diagnoses, which are not included in the research data (Herrett *et al*, 2015). Moreover, binary classifications – e.g. person has a disease or not – do not reflect the continuum of health that most people experience, and risk factors are measured much more frequently among those in 'at risk' categories (e.g. those with diabetes) than the general population (Herrett *et al*, 2015). For income data, not all income is subject to taxation, and hence may not appear on the administrative data. Clergy in the U.S. may set aside up to 50% of their pay as a housing allowance (to offset those who serve congregations with a rectory or parish house). The earnings reported to the government are thus lower than the actual payments made to the individuals and have been observed in comparing survey and administrative records (Bollinger *et al*, 2018). Hand (2018) notes that "A particular issue with administrative data sets arises from the very fact that they were not deliberately collected to answer the later statistical question being addressed." He cites the example of UK crime statistics based on surveys and administrative records, which often trend in opposite directions. He rejects the idea that the latter must be more accurate, as administrative data must be taken as they are while the survey questions are now based on much research into how to elicit the desired information.

While measurement error is believed to be larger in self-report surveys, administrative data are not without r*eporting error*. Administrative data are often a compilation of data from different sources. Administrative data can be generated by computerized processes, such as automated payments of pensions or State benefits, where the recipient, date, value, and type of payment are automatically recorded as part of the payment process. Such processes might be centralised, for example administered by one central government administrative unit, or decentralised, for example administered in parallel by regional offices. If generated by a centralised process, such data are often considered reliable and accurate (see for example Jacobebbinghaus and Seth, 2007). With decentralised processes the data captured can be inconsistent, for example due to differences in computing systems or reporting requirements. Administrative data can also be generated by individuals filling in paper or online forms to report information on themselves (such as tax forms or applications for State benefits). Finally, administrative data can be generated by third parties reporting about individuals, for example health administrators, employers, or schools filling in forms to fulfil government reporting requirements about patients, employees, or pupils. There is no compelling reason to think that administrative data generated by individuals or third parties filling out forms are less subject to reporting error than survey data.

One example is reporting of race on death certificates in the U.S., which is often based on observation by the medical examiner rather than on self-report from family members of the deceased. Similarly, on UK death certificates occupation is recorded by a family member who may not describe a job in the same way as the deceased would have, creating errors when looking at mortality by occupational class (Townsend *et al*, 1992). At the other end of the life course, Connelly and Gayle (2017) argue that parental occupational information from administrative birth records should not be assumed to be suitable for analysis. Administrative data on earnings collected for taxation purposes may be biased downward since individuals have incentives to hide earnings (see Bollinger *et al*, 2018).

The quality of measurement of administrative data also depends on whether the data are central to the administrative purpose or not. For example, in data about social security entitlement reported to the State by employers, information about earnings and tax deductions are likely to be subject to more quality control than information about the employee's level of education, which is not critical to the calculation of social security entitlements. If administrative data are associated with performance management, then it can be susceptible to distortion. A much-discussed instance of this is educational testing. These tests can be "high stakes": in different contexts standardized educational tests may significantly affect outcomes for students (for example grade promotion), teachers (employability or bonus pay) or schools (reputation, resources). This can lead to changes in the behaviour of individuals involved in the data generating process. Measurement error in these cases would include cheating by students, teachers or administrators. (Coverage error may be induced by the strategic exclusion of low-performing students from taking the test.) For further discussion see Nichols and Berliner (2005). More generally, as administrative data are typically more directly related to resource allocations and social decisions, there may be, in some contexts, more incentive for individuals involved in the data generating process to ensure accuracy of the resulting data than in surveys, but in other cases the incentives may go the other way. For example, the introduction of the Quality and Outcomes Framework in the UK, which provides incentives to GPs to identify and accurately code some kinds of activities, e.g. smoking behaviours, more than others, can distort information on the prevalence of risk factors since they are records more in response to the availability of fees that population behaviours (Herrett *et al*, 2015).

Finally, *processing error* may occur in administrative data, especially in the transfer of paper records to electronic formats. Descriptions may be coded incorrectly; transcription errors may occur and other errors may arise through the process of transforming administrative forms to analytic data.

Errors Entering Through Data Integration

Starting with errors of representation, coverage may be improved by integration, particularly when the enhanced data includes the union of the units in the survey and administrative sources. However, when data are linked to enhance measurement or content, *coverage errors* may be introduced or exacerbated. At best, the coverage of the linked data will be the intersection of the coverage of the component data sets: and this will often be less than the coverage of either. That is, gaps in one data source may be compounded with those in the other.

Where data are being linked, further errors of representation can arise through *failure to link*, or, when linkage requires consent, through *failure to obtain consent* (Sakshaug and Antoni, 2017). With probabilistic matching (without unique identifiers) there will be both mis-matches and failures to make correct matches. Even with deterministic matching (using a unique identifier) there can be errors, if the unique identifiers are mis-recorded in one or more source data sets. This can happen, for example if identifiers are reported in a survey or to an administrative organization.

Failure to obtain consent or to link will introduce selection bias if such failures are related to unit characteristics. Existing evidence confirms that matching errors are not likely to be random, but rather related to the characteristics of the units under study. Sakshaug *et al* (2017) examine the rate of linkage errors using different procedures to link a German federal employment database to a general population survey. Identifiers on the sampling frame are error-prone and non-unique. They report linkage rates of between 60 and 80% and find that the linkage rate, and hence linkage error, varies with individual and household characteristics. See also Sakshaug and Antoni (2017) and Aldridge *et al* (2015). Bollinger *et al* (2019) find that linkage rates between the U.S. *Current Population Survey* and federal tax data are over 90% for most groups but fall to 50% for Hispanic non-citizens (they are well over 80% for Hispanic citizens). Illegal immigrants are included in the survey, but not in the administrative tax records.

Various kinds of measurement error can also be introduced through the process of data integration. Specification errors may arise through differences in units of aggregation between the two data sources, producing many-to-one or one-to-many problems. Similarly, specification errors may also be introduced though temporal or periodicity mismatches. Survey data are typically designed to collect information about a specific point in time or period. Administrative data are continuously updated and may refer to administrative units, e.g. a hospital episode, rather than a time period. Specification errors arise through mismatches in the time period covered by the two data sources, or delays in updating administrative data. For example, survey data collection on employment and earnings often refer to the past pay period, past month or calendar year, while administrative records may refer to a tax year.

Decisions about how to resolve conceptual differences may introduce measurement error in the integrated data set. If similar variables occur in both data sources, decisions about which one to use or how to combine the two may produce errors.

Kapteyn and Ypma (2007) examine measurement error in a survey conducted in Sweden which was linked to the *Longitudinal Individual Data* (LINDA). They estimate a set of models examining difference between survey and administrative measures of earnings, pensions and taxes allowing for mis-linkage between the survey and the administrative records of LINDA. They estimate the mis-linkage rate to be between 0.02 and 0.13. They also conclude that estimates of measurement error in survey data which do not allow for linkage mismatch may lead to biased estimates of the measurement error process. Bollinger *et al* (2018) also estimate similar models, but find less evidence of linkage error.

4. Conclusion

Integrated data is becoming increasingly common. It has been used to some good effect in a variety of contexts, although predominantly in validation studies where the primary goal was to investigate the quality of survey data. Increasingly, integrated data is being used in an enhancement role, either to improve representation, content or measurement. However, these limited applications are just the beginning of what may be possible with integrated data.

While the potential for integrated data is great, little has been developed as standards for the use of integrated data. Researchers need to pay close attention to issues of representation, measurement and specification. In the past, researchers have been content to treat administrative data sources as a gold standard; however recent evidence suggests a more comprehensive view is wiser. The processes which generate administrative data are very different than those of surveys, but they are not without measurement error, sample selection and other data errors. The process of linking or appending data can itself generate significant issues which are largely ignored. While integrated data can solve many problems, the solution is not costless and not always simple.

Researchers need a systematic approach to integrated data, which we argue should be based on the Total Error framework. Assessment of data quality at every stage and from every source is an important step in all research projects. While there are many standards for such approaches in different fields, none of these standards has addressed the issues of integrated data in a comprehensive way.

The need for a systematic approach suggests a research agenda on methodology seeking to provide researchers with both broad guidelines and specific approaches which can be implemented across many fields in social sciences for many different types of integration. Drawing on the Total Error framework we think three basic areas should be particular priorities for initial investigation:

1. Errors of representation created in the linkage process
2. Best practice for validation studies

3. Specification errors arising in enhancement applications

Link failure, link mismatch, coverage universe of administrative data and sampling universe of survey data are all potentially present in any linkage of survey and administrative data. Using the TE framework allows investigation of the trade-off between reducing omitted variable bias and potentially increased representation error in the case of enhancement data. This includes the trade-off between having the full population in an administrative data source (unlinked) but missing many key variables. By establishing a methodology which is well suited to measuring the costs and benefits of any of the approaches, this allows cohesive investigation of potential solutions or improvements such as inverse probability weighting.

Validation type studies are fundamentally a repeated measures opportunity. Historically, integrated data which includes two measures of the same variable has treated the administrative data as the gold standard. There are times when this assumption may be justified; however, few guidelines exist for this decision. A testing environment is desirable and the TE framework can be used to develop one, again measuring trade-offs between bias and specification. There is little guide on how to address and handle cases where neither survey nor administrative data can be treated as a gold standard.

Specification errors arise in particular when surveys and administrative data differ in their time frame or unit of analysis. There are few if any comprehensive approaches to dealing with this issue which will be increasingly more common. Again, the Total Error framework allows an investigation of the trade-offs between enhancement or validation data advantages in reducing other biases and the cost of misspecification in this dimension.

References

Abowd, J. M. and M. H. Stinson (2013) “Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data.” *Review of Economics and Statistics*, XCV(5):1451-67.

Aldridge, R. W., Shaji, K., Hayward, A. C. and L. Abubakar (2015) “Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies.” *PLOS One* https://doi.org/10.1371/journal.pone.0136179.

Baron, J. D., and D. A. Cobb-Clark. (2010) *Are Young People’s Educational Outcomes Linked to Their Sense of Control?* SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Battistin, E., DeNadai, M. and B. Sianesi (2014) “Misreported Schooling, Multiple Measures and Returns to Educational Qualifications.” *Journal of Econometrics*, 181: 136-150.

Biemer, P.P. (2014) “Total (Survey) Error: Adapting the Paradigm for Big Data.” Paper presented at the International Total Survey Error Workshop, Washington, DC, October.

Biemer, P.P., Eckman, S., Edwards, B., de Leeuw, E., Kreuter, F., Lyberg, L., Tucker, C., and B. West (eds.) (2017), *Total Survey Error in Practice*. New York: Wiley.

Bollinger, C.R. (1998) “Measurement Error in the CPS: A Nonparametric Look.” *Journal of Labor Economics*, 16(3):576-94.

Bollinger, C.R. and M.H. David, (1997) “Modeling Discrete Choice with Response Error: Food Stamp Participation.” *Journal of the American Statistical Association*, 92(439):827-35.

Bollinger, C.R., Hirsch, B.T., Hokayem, C. and J.P. Ziliak (2018) *The Good, The Bad, and The Ugly: Measurement Error, Non-response and Administrative Mismatch in the CPS,* http://christopherbollinger.com/.

Bollinger, C.R., Hirsch, B.T., Hokayem, C., and J.P. Ziliak (2019) “Trouble in the Tails? What we know about earnings nonresponse thirty years after Lillard, Smith, and Welch.” *Journal of Political Economy,* 127(5):2143-85.

Bound, J. and A.B. Krueger (1992) "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics,* 9:1-24.

Bound, J., Brown, C., Duncan, G.J. and W.L. Rodgers, (1994) "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data." *Journal of Labor Economics*, 12(3):345-68.

Bozio, A., Crawford, R., Emmerson, C. and G. Tetlow (2010) *Retirement Outcomes and Lifetime Earnings: Descriptive Evidence from Linked ELSA-NI data* Department for Work and Pensions Working Paper No. 81, Norwich: HMSO.

Bozio, A., Emmerson, C., O'Dea, C. and G. Tetlow (2013) *Savings and Wealth of the Lifetime Rich: Evidence from the UK and US* IFS Working Papers. London: Institute for Fiscal Studies.

Bozio, A., Emmerson, C., O'Dea, C. and G. Tetlow (2017) "Do the Rich Save More? Evidence from Linked Survey and Administrative Data." *Oxford Economic Papers,* 69(4): 1101–19.

Broecke, S. (2012), "University Selectivity and Earnings: Evidence from UK Data on Applications and Admissions to University." *Economics of Education Review* 31(3): 96–107.

Browning, M. and S. Leth-Petersen (2003) "Imputing consumption from income and wealth information." *Economic Journal*, 113(488): F282-F301.

Burkhauser, R. V., Hérault, N., Jenkins, S.P. and R. Wilkins (2017) *Survey Under-Coverage of Top Incomes and Estimation of Inequality: What is the Role of the UK's SPI Adjustment?* NBER Working Papers 23539, Cambridge MA: National Bureau of Economic Research, Inc.

Burström, B. and P. Fredlund (2001) "Self rated health: Is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes?" *Journal of Epidemiology and Community Health* 55:836–40.

Cajner, T., Crane, L. , Decker, R. , Hamins-Puertolas, A. and C. Kurz, (2019), *Improving the Accuracy of Economic Measurement with Multiple Data Sources: The Case of Payroll Employment Data*, Finance and Economics Discussion Series 2019-065. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2019.065

Card, D., Hildreth A.K.G. and L.D. Shore-Sheppard (2004) "The Measurement of Medicaid Coverage in the SIPP: Evidence from a Comparison of Matched Records." *Journal of Business and Economic Statistics*, 22(4):410-20.

Cobb-Clark, D.A., Kassenboehmer, S.C. Le, T., McVicar, D. and R. Zhang (2015) " 'High'-School: The Relationship between Early Marijuana Use and Educational Outcomes." *Economic Record,* 91(293): 247–66.

Cohen, S.B. and B.L. Carlson (1994) "A Comparison of Household and Medical Provider Reported Expenditures in the 1987 NMES." *Journal of Official Statistics*, 10(1):3-29.

Connelly, R. and V. Gayle (2017) "An Investigation of the Consistency of Parental Occupational Information in UK Birth Records and a National Social Survey." *European Sociological Review* 33(2): 240–56.

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

Cornish, R.P., Tilling, K., Boyd, A., Davies A. and J. Macleod (2015) "Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years." *International Journal of Epidemiology*, 44(3): 937–45.

Crawford, R., and C. O'Dea. (2014) *Cash and Pensions: Have the Elderly in England Saved Optimally for Retirement?* IFS Working Papers. London: Institute for Fiscal Studies.

Crawford, R., Keynes, S. and G. Tetlow (2014), *From Me to You? How the UK State Pension System Redistributes.* IFS Working Papers, London: Institute for Fiscal Studies.

Dale, S.B. and A.B. Krueger (2014) "Estimating the Effects of College Characteristics over the Career Using Administrative Earnings Data." *Journal of Human Resources,* 49(2): 323–58.

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

Department for Work and Pensions (2015) *Households Below Average Income An Analysis of the Income Distribution 1994/95–2013/14*. London: Department for Work and Pensions.

Dearden, L., Miranda, A. and S. Rabe-Hesketh (2011) "Measuring School Value Added with Administrative Data: The Problem of Missing Variables." *Fiscal Studies*, 32(2): 263–78.

Doll, R. and A.B. Hill (1954) “The mortality of doctors in relation to their smoking habits; a preliminary report.” *British Medical Journal*, 1(4877): 1451–55.

Falster, M.O., Leyland A.H. and L.R. Jorm (2019) “Do hospitals influence geographic variation in admission for preventable hospitalisation? A data linkage study in New South Wales, Australia.” *BMJ Open*, 9(2):e027639.

Geue, C., Lorgelly, P., Lewsey, J., Hart, C. and A. Briggs (2015) “Hospital Expenditure at the End-of-Life: What Are the Impacts of Health Status and Health Risks?” *PLoS One* 10(3): e0119035.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L., Smith, P., Dibben C. and H. Goldstein (2018), “GUILD: GUidance for Information about Linking Data sets.” *Journal of Public Health*, 40(1):191–98.

Goudreau, K., Oberheu, H., and D. Vaughan (1984) “An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children Program.” *Journal of Business and Economic Statistics*, 2(2): 179-186.

Gray, L., Gorman, E., White, I.R., Katikireddi, S.V., McCartney, G., Rutherford, L., and A.H. Leyland (2019) “Correcting for non-participation bias in health surveys using record-linkage, synthetic observations and pattern mixture modelling.” *Statistical Methods in Medical Research* https://doi.org/10.1177/0962280219854482

Griffiths, l.J, Lyons, R.A, Bandyopadhyay, A, Tingay, K.S, Walton, S., Cortina-Borja, M., Akbari, A., Bedford, H. and C. Dezateux (2017) “Childhood asthma prevalence: cross-sectional record linkage study comparing parent-reported wheeze with general practitioner-recorded asthma diagnoses from primary care electronic health records in Wales”. *BMJ Open Respiratory Research*, 5(1): e000260.

Griffiths, L.J., Cortina-Borja, M., Bandyopadhyay, A., Tingay, K., De Stavola, B.L., Bedford, H., Akbari, A., Firman, N., Lyons, R.A. and C. Dezateux (2019) “Are children with clinical obesity at increased risk of inpatient hospital admissions? An analysis using linked electronic health records in the UK Millennium Cohort Study.” *Pediatric Obesity* 14(6): e12505.

Groen, J.A. (2012) “Sources of error in survey and administrative data: the importance of reporting procedures.” *Journal of Official Statistics*, 28(2):173-98.

Groves, R.M. (1989) *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. (2011) “Three Eras of Survey Research.” *Public Opinion Quarterly*, 75(5): 861-71.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and R. Tourangeau (2009) *Survey Methodology* (2nd ed.), Hoboken, NJ: Wiley.

Haan, M., Kaplan, G.A. and T. Camacho (1987) “Poverty and Health Prospective Evidence from the Alameda County Study.” *American Journal of Epidemiology* 125(6): 989-98.

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

Harron, K.L., Doidge, J.C., Knight, H.E., Gilbert, R.E., Goldstein, H., Cromwell, D.A. and J.H. van der Meulen, (2017) “A guide to evaluating linkage quality for the analysis of linked data.” *International Journal of Epidemiology*, 46(5):1699–1710.

Hart, C.L., MacKinnon, P.L., Watt, G.C., Upton, M.N., McConnachie, A., Hole, D.J., Smith, G.D., Gillis, C.R. and V.M. Hawthorne (2005) “The Midspan studies.” *International Journal of Epidemiology* 34(1): 28-34.

Herrett, E., Gallagher, A. M., Bhaskaran, K., Forbes, H., Mathur, R., van Staa, T., and L. Smeeth (2015) “Data Resource Profile: Clinical Practice Research Datalink (CPRD).” *International Journal of Epidemiology*, 44(3), 827–36.

Jäckle, A. and S. Eckman (2019) “Is that still the same? Has that Changed? On the Accuracy of Measuring Change with Dependent Interviewing.” *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smz021.

Jacobebbinghaus, P. and S. Seth (2007) “The German integrated employment biographies sample IEBS.” *Schmollers Jahrbuch: Journal of Applied Social Science Studies,* 127(2):335-42.

Jenkins, S.P. (2017) “Pareto Models, Top Incomes and Recent Trends in UK Income Inequality.” *Economica*, vol. 84(334):261-89.

Juhn, C., and K. McCue (2016) “Selection and Specialization in the Evolution of Marriage Earnings Gaps.” *RSF: The Russell Sage Foundation Journal of the Social Sciences,* 2(4):237–69.

Kapteyn, A. and J.Y. Ypma (2007) “Measurement Error and Misclassification: A Comparison of Survey and Administrative Data.” *Journal of Labor Economics*, 25:513-51.

Klein, B. and D. Vaughan (1980) “Validity on AFDC Reporting Among List Frame Recipients.” *Reports from the Site Research Test*, ed. Janice Olson, Washington, D.C.: U.S. Department of Health and Human Services, Assistant Secretary for Planning and Evaluation, Chapter 11.

Kreuter, F., Presser, S. and R. Tourangeau (2008) “Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity.” *Public Opinion Quarterly*, 72(5): 847–865.

Kreuter, F., G. Muller and M. Trappmann (2010) “Nonresponse and Measurement Error in Employment Research.” *Public Opinion Quarterly*, 74(5):880-906.

Lavrakas, P.J. (2013) “Applying a Total Error Perspective for Improving Research Quality in the Social, Behavioral, and Marketing Sciences.” *Public Opinion Quarterly*, 77(3): 831-50.

Lynn, P., Jäckle, A., Jenkins, S.P. and E. Sala (2012) “The Impact of Interviewing Method on Measurement Error in Panel Survey Measures of Benefit Receipt: Evidence from a Validation Study.” *Journal of the Royal Statistical Society, Series A,* 175(1):289-308.

Marmot, M.G. and M.J. Shipley (1996) “Do socioeconomic differences in mortality persist after retirement? 25 Year follow up of civil servants from the first Whitehall study.” *British Medical Journal* 313:7066: 1177-80.

Mathiowetz, N.A. and G.J. Duncan (1988) “Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment.” *Journal of Business and Economic Statistics*, 6(2):221-229.

Marquis, K.H., and J.C. Moore (1990), “Measurement errors in survey of income and program participation (SIPP) program reports.” *Proceedings of the Sixth Annual Research Conference (U S Bureau of the Census, Washington, DC)*, 721-45.

Mellow, W. and H. Sider (1983) “Accuracy of Response in Labor Market Survey: Evidence and Implications.” *Journal of Labor Economics*, 1(4):331-34.

Meyer, B.D., Wallace, K.C., Mok, K.C. and J.X. Sullivan, (2015) “Household Surveys in Crisis.” *Journal of Economic Perspectives*, 29(4):199-226.

Meyer, B. D., Mittag, N. and R. George (2018) “Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation.” *NBER working paper 25143*.

Moore, J.C. and K.H. Marquis (1989) “Using Administrative Record Data to Evaluate the Quality of Survey Estimates.” *Survey Methodology*, 15(1):129-43.

Morgan, J.N. (1989) *Panel Study of Income Dynamics, 1968-1987: Validation Study*, Survey Research Center, University of Michigan.

Nichols, S. and D. Berliner, (2005) *The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing*. Education Policy Research Unit Report: Arizona State University.

Overbeek A, van den Berg M.H., Hukkelhoven C.W., Kremer L.C., van den Heuvel-Eibrink M.M., Tissing W.J., Loonen J.J., Versluys A.B., Bresters D., Kaspers G.J., Lambalk C.B., van Leeuwen F.E., van Dulmen-den Broeder E.; and DCOG LATER/VEVO Study Group. (2013) “Validity of self-reported data on pregnancies for childhood cancer survivors: a comparison with data from a nationwide population-based registry.” *Human Reproduction*; 28(3):819-27.

Pascale, J., Roemer, M.I. and D.M. Resnick (2009) “Medicaid Underreporting the CPS: Results from a record check study.” *Public Opinion Quarterly*, 73(3):497-520.

Rodgers, W.L. and A.R. Herzog (1987) “Covariances of Measurement Errors in Survey Responses.” *Journal of Official Statistics*, 3(4):403-18.

Rogers, W.L., Brown, C. and G.J. Duncan, (1993) “Errors in Survey Reports of Earnings, Hours Worked and Hourly Wages.” *Journal of the American Statistical Association*, 88(424):1208-18.

Roemer, M. (2002) *Using Administrative Earnings Records to Assess Wage Date Quality in the March Current Population Survey and the Survey of Income and Program Participation*, LEHD Technical Paper TP-2002-22, Washington DC: US Census Bureau.

Rouse, C.E., Hannaway, J., Goldhaber, D. and D. Figlio, (2013) “Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure.” *American Economic Journal: Economic Policy*, 5(2):251–81.

Sakshaug, J.W. and M. Antoni (2017) “Errors in Linking Survey and Administrative Data.” In *Total Survey Error in Practice* (eds P.P. Biemer, E. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker and B.T. West), 557-74.

Sakshaug, J.W. and M. Antoni (2019) “Evaluating the Utility of Indirectly Linked Federal Administrative Records for Nonresponse Bias Adjustment.” *Journal of Survey Statistics and Methodology*, 7(2):227–49.

Sakshaug, J.W., Antoni, M. and R. Sauckel (2017) “The Quality and Selectivity of Linking Federal Administrative Records to Respondents and Nonrespondents in a General Population Sample Survey of Germany.” *Survey Research Methods*, 11(1), 63-80.

Stinebrickner, T.R. and R. Stinebrickner (2003) “Understanding Educational Outcomes of Students from Low Income Families: Evidence from a Liberal Arts College with a Full Tuition Subsidy Program.” *Journal of Human Resources*, 38(3):591-617.

Stinebrickner, T.R. and R. Stinebrickner (2008) “The Effect of Credit Constraints on the College Drop-Out Decision: A Direct Approach Using a New Panel Study.” *American Economic Review*. 98(5):2163-84.

Stinebrickner, T.R. and R. Stinebrickner (2012), “Learning about Academic Ability and the College Drop-Out Decision.” *Journal of Labor Economics,* 30(4):707-48.

Townsend, P., Whitehead, M. and N. Davidson (1992) *Inequalities in health: The Black Report and the Health Divide,* London: Penguin Books.

UK Statistics Authority (2014) *Quality Assurance and Audit Arrangements for Administrative Data*. London: UK Statistics Authority.

Vaughan, D. (1978) “Errors in Reporting Supplemental Security Income Recipiency in a Pilot Household Survey.” *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, 288-93.

Vaughan, D. and R. Yuskavage (1976) “Investigating Discrepancies between Social Security Administration and Current Population Survey Benefit Data for 1972.” *Proceedings of the Social Statistics Section, American Statistical Association.* Part 2: 824-9.

Vaughan, D., Lininger, C. and R. Klein (1983) “Differentiating Veterans’ Pensions and Compensation in the 1979 ISDP Panel.” *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Whitley, E., Batty, G.D., Hunt, K., Popham, F. and M. Benzeval (2014) “The role of health behaviours across the life course in the socioeconomic patterning of all-cause mortality: the west of Scotland twenty-07 prospective cohort study.” *Annals of Behavioral Medicine* 47:2: 148-57.