Deliverable Number: D6.6

Deliverable Title: Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of administrative data

Work Package: 6: New forms of data: legal, ethical and quality matters

Deliverable type: Report

Dissemination status: Public

Submitted by: NIDI

Authors: George Groenewold, Susana Cabaco, Linn-Merethe Rød, Tom Emery

Date submitted: 23/08/19

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 654221.

www.seriss.eu @SERISS_EU

SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe's grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey for Health Aging and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: Groenewold, G., Cabaco, S., Rød, L.M., Emery, T., (2019) Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of administrative data. Deliverable 6.6 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. Available at: wwwww.seriss.eu/resources/deliverables
www.seriss.eu/resources/deliverables

Content

1 Introduction

Administrative data sources are increasingly being made available for research purposes. The increased use of administrative data sources as a basis for survey sampling methodologies widen the scope and capacity for linking administrative data and survey data for research purposes. Furthermore, the potential for such linkage is enhanced by the bridging of scientific and civil service data infrastructures and the deeper integration of science and the policy making process. By linking data, policy makers can validate their data with scientifically driven measures and social scientists can better engage in questions with direct relevance for policy makers.

This potential is not easily realised however. The data infrastructures of administrative data sources and social surveys are generally governed by different laws, different ethical practices and different priorities and purposes. Operating across these differences poses several challenges.

Building on task 6.2 and Work package 6 in the SERISS project, this report addresses the legal requirements and ethical challenges in linking administrative data and survey data. This document is the output of deliverable 6.6 that aims to provide an overview of administrative data usage in Social Science Research in Europe, and specifically in survey research.

Work Package 6 of the SERISS project addresses the major legal and ethical challenges facing cross-national social science research which relies on access to large-scale data on an individual level. The focus is on social surveys and the use of new data types in a social survey context in particular, including biomarker, social media data and administrative data.

The focus of Task 6.2 was the legal requirements and ethical challenges that may come about when survey data are linked with administrative data sources. The task addresses the issues that need to be taken to meet these challenges in order to increase and improve the research use of these data sources[1]. The main purpose of this report was to provide an overview of the legal and ethical issues in, and to overcome possible barriers to, research use of administrative data. As such this deliverable is aimed at researchers and infrastructures managing new forms of data.

Administrative data can be defined as the “information collected primarily for administrative purposes […] collected by government departments and other organizations for registration, transactions and record keeping”[2]. The value of micro-level administrative data for social science empirical research and public policy evaluation has been extensively recognized in different sectors, given the rich demographic and socio-economic information recorded (Connelly et al., 2016; Card et al., 2010).

More specifically, there are crucial differences between survey microdata and administrative data, well summarized by Card et al. (2010). First, administrative data covers (significantly)

[1] European Commission, Directorate-General for Research and Innovation (2016: 45).

[2] See https://adrn.ac.uk/for-the-public/faq/about-the-data/.

larger proportions of the population – in some instance's full population records[3] (as in population registers) - when compared to most social surveys. It is easily apparent why larger samples are very attractive to researchers: a good example could be the study of factors associated with educational attainment among specific social groups (e.g. ethnic minorities) in small geographical areas or instances of rare behaviors or population characteristics.

Secondly, most administrative data have a longitudinal structure that allows researchers to include a time dimension in their analysis[4] and, for example, track long-term effects of certain events (illness, job loss, etc.) or evaluate the pre and post policy implementation contexts in a given area. These authors also argue that administrative data provide higher quality information when compared to most survey sources (which tend to be more prone to non-response, attrition and other types of measurement bias), covering also areas that are not included in social surveys (Card et al., 2010). However, administrative data sources are not primarily collected or oriented towards specific scientific research purposes and might even require some data preparation (e.g. constructing and recoding variables) before the analysis.

The use of administrative data linked to surveys in social science research, despite not being widespread, has been increasingly recognized as a good research practice (Card et al., 2010; Connelly et al., 2016; Poulain and Herm, 2013). Several researchers and organizations have indeed advocated for the development and expansion of access to administrative microdata, given that it is "critical for cutting-edge empirical research" (Card et al., 2010: 1). In particular, when administrative data is linked to survey data, there is the potential added benefit of data validation, i.e. the researcher can use the survey data in order to assess the accuracy of the administrative data records.

Furthermore, there is another important gain derived from the usage of administrative data associated with the fact that part of the information that social scientists are interested in measuring is already captured by administrative records. The usage of administrative data can then contribute to reduce the burden on survey respondents which represents an important goal in social science research, not only due to the rising costs of survey data collection, but also because administrative records can potentially provide a more comprehensive picture in certain domains (for example, tax records or social security payments data).

These benefits of administrative data linkage are well known, widely accepted and there is broad consensus that administrative data linkage should be encouraged to facilitate better science. However, there are several obstacles and barriers to administrative data linkage. Foremost of these are statistical disclosure concerns and data privacy legislation which have been at the forefront of the arguments advanced by data owners/custodians – typically national statistical offices – to restrict and limit access to administrative data. The strictness of data protection legislation varies by country, which creates additional challenges to

[3] The collection of data has a legal basis and residents must ensure that certain personal data are recorded as required by government-controlled administrative systems.

[4] On what concerns population registers, Poulain and Herm (2013: 202) note: "A central register makes it possible to bring together all the events involving a given individual, exhaustively for the entire population".

programs designed to provide access to administrative data from different countries (e.g. Data without Borders of the International Data Access Network).

In recent years however there have been several national and European legislative initiatives to improve access to administrative data. At the international level, the General Data Protection Regulation (GDPR), EU Regulation 2016/679 focuses on the protection of personal data[5] and the free movement of such data[6], replacing the EU Data Protection Directive (95/46/EC). The primary intention of GDPR is to harmonize data protection law across the EU. GDPR places more obligations on the data controller and processor than the former Directive 95/46/EC – the Regulation lists the overall principles relating to any handling of personal data, in which the controller is responsible for, and must be able to demonstrate compliance with those principles. The GDPR also includes important exemptions from purpose limitation - personal data collected for one purpose should not be used for a new (incompatible) purpose - for research and archiving of personal data, granting that scientific research shall not be considered to be incompatible with the initial purposes ('purpose limitation'). This creates the conditions for the reuse of personal data for scientific research, if the appropriate data protection safeguards are in place.

These developments have also been mirrored within the scientific community by the widespread adoption of the FAIR principles for data access[7]. These principles outline how data should be made findable, accessible, interoperable and reusable. This is of considerable relevance for administrative data in the social sciences as it could be broadly said to fail on all counts. There are significant impediments on all four points and the development of FAIR data infrastructure within the context of the European Science Cloud could be an opportunity to rapidly increase the findability, accessibility, interoperability and reusability of administrative data for the social sciences and beyond.

In addition to the legal facilitation of access to administrative data for research purposes and pursuit of the FAIR principles, there have also been several technical developments which further enable secure administrative data access and can potentially provide new and secure ways to access the highly sensitive yet scientifically pertinent data that administrative records contain. The aim of this report is to provide an overview of these legal, ethical and technical developments in administrative data linkage and evaluate potential future frameworks for such linkage in the context of the European Research Infrastructure landscape.

The report begins with a general overview of research that has utilized administrative data for research purposes and examines the different frameworks used for accessing, linking,

---

[5] Personal data is defined as "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (European Parliament and Council Regulation 2016/ 679: 33).

[6] The GDPR also defines the terms in which the data should be processed: "The right to protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights" (European Parliament and Council Regulation 2016/ 679).

[7] https://www.go-fair.org/fair-principles/

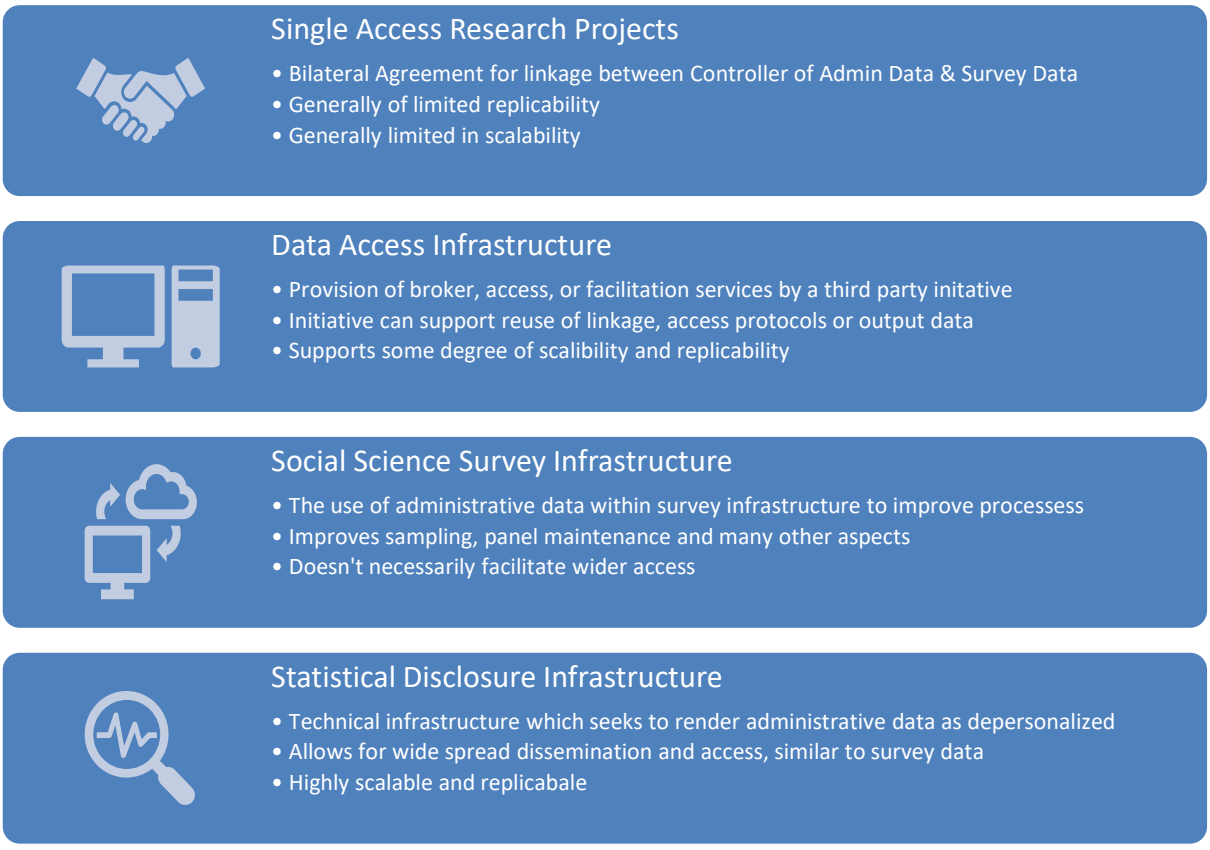
analyzing and disseminating administrative data. The third section of the report reflects on the legal and ethical framework of administrative data linkage and how it affects these infrastructures for administrative data usage. The report concludes with a brief evaluation of the administrative data linkage infrastructure and possible avenues of further development and exploitation.

2 Frameworks for Administrative Data Use

This section introduces diverse studies, initiatives and data infrastructures that have been using or contributing to advance the usage of administrative data in social science research. These are clustered into four groups which attempt to support administrative data through different access models. We begin with the most common which is the administrative data linkage of single access research projects. These projects link survey data and administrative data on a case by case, with little continuity in access or linkage protocols. These tend to be a bilateral agreement and collaboration between the administrative data provider and the researcher.

Some initiatives have tried to extend this and make such administrative data linkage more systematic, predictable and transparent through the development of dedicated administrative data access initiatives. These initiatives are increasingly commonplace but the amount of research they support is still relatively limited. We examine what research has been conducted using such initiatives and the lessons learned. We then examine two forms of administrative data access initiatives which are relevant in the context of survey research. The first is administrative data access through statistical disclosure controls. This allows for data to be accessed more readily but limits the functionality of such data for survey research.

In contrast, there are initiatives advanced by social science infrastructures, which integrate administrative data in survey processes themselves (e.g. Swedish Generations and Gender Survey). Such linkage has very large potential for supporting the improvement of Social Science Survey Infrastructures in Europe through improved sampling, stronger panel maintenance and improved measures of social phenomenon.



These four frameworks for administrative data linkage are exceptionally varied and entail specific legal, ethical and technical considerations. However, to make use of the substantial

scientific potential captured within administrative data, it is necessary that all types of access are supported.

Single access research projects

Single access research projects are designed, prepared and implemented by an individual researcher or research team who holds linkable survey data for which they are data controller. This type of research project usually requires accreditation with the administrative data controller who determines whether the research is ethical, lawful, has scientific merit and has a potential benefit for society. Data controllers only supply the data collections relevant to the research project and require that the researcher using the data preserves the confidentiality of any personal data and agrees not to use the data to attempt to obtain or derive any personal information. Given this, the linkage that is conducted is limited to either a single research project or to a single linkage with a survey infrastructure. Each project therefore has limited interactions or spill over effects for other linkage projects.

Such projects often, though not always, share certain characteristics:

- Administrative data controllers are often providing access to such projects as a secondary rather than primary function of their work.
- They are often conducted ad-hoc, leaving little protocol or procedure in place for subsequent linkage projects of a similar nature;
- They are often post-hoc in that linkage is conducted after the survey data has been collected;
- They are often bilateral between a researcher who holds survey data and an administrative data controller.

Examples of Research

Linking administrative data and survey data offers the potential for many new research opportunities for scientific and policy-related projects. From an assessment of the academic literature, the number of studies based on linkage of records of data systems has been growing, but the actual number of studies remains small relative to the field overall, particularly in the social sciences. Such studies are far more common in public health research, as illustrated by the studies reported in the annotated bibliography in Annex 1. Nevertheless, most existing examples of research which combine administrative data with survey data are derived from single access research projects. Here, we provide a couple of examples of administrative data linkage that have been conducted and illustrate the scientific potential of such linkage. This list is by no means exhaustive and merely serves as an indication of the types of research administrative data linkage currently facilitates.

The field of economics have seen relatively high activity with regards to administrative data linkage, even if administrative data is used in only a small fraction of published papers overall. This high prevalence relative to other fields is primarily due to the nature of administrative data and the concepts it captures. The concepts measured within economics are generally tangible, measurable and captured through administrative processes (e.g. prices, income, assets etc.) Good examples of recent research in this area include Abowd, Mckinney and Zhou (2018) who used data from the Social Security Administration in the US and link this to survey data on household incomes to create a longitudinal file on incomes which builds on the strengths of both self-reported and administratively recorded income levels. This allowed them to analyze changes in the income distribution over time in the US which carefully encapsulated vulnerable, small and marginalized populations that are not

well captured by administrative data. One important thing to not here is that the authors are employees at the US Census Bureau. This helps facilitate access to such data.

In Sweden, Karadja, Mollerstrom and Seim (2017) linked data from administrative income records which provide a reliable measure of income with survey data which asked respondents about their perceived position in the income distribution. This illustrates the potential insights from combining the objective and administrative measures available in government data with subjective and sociological data available via surveys. The research was particularly innovative in that the administrative data was linked ex-ante which allowed researchers to then inform survey respondents of their true position in the income distribution and then monitor changes in attitudes and values.

In Germany, Rasner et al. (2013) used administrative pension data (Statutory Pension Insurance and divorce statistics) linked with survey data from the German Socio-Economic Panel in order to calculate the present value of pension entitlements and analyze sources of inequalities between individuals (in particular, between occupational groups). The linkage of this data was conducted by the German Socio-Economic Panel and the results revealed large inequalities in Social Security Wealth between different occupational sectors and between various socio-economic groups.

Another example is Longhi's study combining UK census data and survey data (British Household Panel Survey), analyzing the impact of cultural diversity (measured by the number and size of ethnic groups living in each Local Authority District in England) on individual wages. The study shows that there is a positive association between diversity and wages, but only when cross-section data are used; panel estimation shows no impact of diversity (Longhi, 2013).

The LONGPOP project - Methodologies and Data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers[8] -, is a four-year project coordinated by the Spanish National Research Council, which intends to facilitate research and increase the usage of a number of datasets that are currently potentially underused. The project intends to capitalize on the fact that "over the past decade research teams across Europe have been involved in the development and construction of longitudinal population registers […] opening up avenues for new linkages between different data sources"[9].

One of the innovative features of this project is the implementation of cross-national registry analysis. This type of data linkage has been rare in the past, mainly due to barriers in the access to foreign registry data and other technical difficulties (in particular, problems with matching and data consistency across countries and time). The project has ten participating beneficiaries plus one partner, including universities, research institutes, enterprises and public administration.

To exemplify the research conducted in the context of LONGPOP, the study 'Early life conditions, place and health in Scotland' (U. of Edinburgh) will examine how early health environment affect subsequent health and mortality, using Scottish Longitudinal Study data

[8] For more details see: http://longpop-itn.eu/project/.

[9] LONGPOP webpage: http://longpop-itn.eu/project/.

(includes census data) for the cohort born in 1936 linked with the 1947 Scottish Mental Health Survey. Given that individual addresses are available at most time points (i.e., events), it is feasible to construct a range of geographical variables and examine the effects of constructs such as area-level deprivation, rurality, access to services and green spaces[10].

Nonetheless, there is an important limitation associated with LONGPOP which concerns the fact that there is no emphasis or provisions for the dissemination and sharing of the research data produced – given this limitation, it can be argued also that the overall investment in LONGPOP brings modest returns. Where possible, it would be important to encourage data sharing and to have a data management plan for each research project. This is a common issue in administrative data linkage projects (Playford et al, 2016)

There are also examples of studies linking experimental data with administrative records. For instance, Chetty et al. (2011), in the Project STAR (Student/ Teacher Achievement Ratio), randomly assigned 11,571 students and their teachers to classrooms in their schools - from kindergarten to third grade. For the analysis, the authors used tax returns data from the U.S. Internal Revenue Service, in order to capture the earnings of those students at age 27. The authors were interested in the impacts of classroom environments on long-term outcomes and their initial findings "raise the possibility that differences in school quality perpetuate income inequality" (2011: 1658).

In all these studies, replication is a serious concern, especially as the analytical syntax and code that are commonly shared with scientific analysis are not commonly used within administrative data linkage projects. Given the generally opaque nature of administrative data and its metadata, such documentation is vital for the peer-review process and replication.

Legal, Ethical and Technical Issues

The paucity of linkage studies cannot solely be attributed to technical issues. Over time deterministic and probabilistic record-linkage methods and software have been developed permitting linkage of records with or without a unique person-identifier number (e.g. Groenewold, van Ginneken and Masseria, 2008; Künn, 2015; Lifang G., Baxter, Vickers and Rainsford, 2003; Sayers, Ben-Shlomo, Blom and Steele, 2015).

The literature indicates that cost considerations are as important as linkage and manual verification of quality of linkage involves costly software and personnel. Technically speaking, success of matching and linking records depends also on the kind of disclosure limitation method used to protect confidential information contained in the data, among others (e.g. Itoh and Takano, 2011). Another major factor is the concern and fear of government institutions about the risk of breaching individual privacy laws. Consent of respondents about sharing private information in administrative data is usually not available. Thus, such kind of legal constraints on the use of administrative data seriously limit development of and access to linked data sets. Several studies urge policymakers, who have a lot to gain from the findings of social science research using linked data, to help create a better procedural infrastructure in support of data linkage projects for scientific research (Künn, 2015).

[10] For details on other project see 'ESR calls' http://longpop-itn.eu/.

Contrary to public health scientists, the interest of social scientists in the usage of administrative data is relatively new. To the extent that demographers and statisticians belong to this group of scientist, these persons have traditionally been involved in the design of national administrative data collection systems (United Nations Department of Economic and Social Affairs, 2016). In the past decades, they also developed a toolbox of methods to assess and monitor the quality, completeness of coverage, and comparability over time and space of administrative data (e.g. Brass, 1975; Hill, 1987). Nevertheless, the usage of administrative data sources is reflected in only a few publications based on administrative data, or administrative data linked to survey data. This is illustrated in relatively few social science publications in Annex 1. These publications are the result of Endnote X6 searches of open-access library catalogue systems around the world, such as Web of Science (SSCI and TS), University of Michigan, Library of Congress, PubMed).

The decision was made to use a selected set of keywords and, in particular, combination of keywords which also include one of the following: administrative data, administrative records, survey, survey data, population registers, population records, small area estimation, data protection, privacy, administrative data linking, linkage, linkage consent, data quality, data coverage, data comparability. In view of the limited number of social science publications derived from literature searches, the scope was extended to also include public health research publications, and reports of studies outside the European context.

The publications in Annex 1 address *obstacles* to the full usage of administrative data sources in social science research, and how to overcome them, including (1) variation in EU national data protection and privacy laws vis-à-vis EU data protection directives, and views of EU citizens regarding linkage of their personal data (i.e. data consent) (e.g. Aldhouse, 2013, 2014; de Hert and Papakonstantinou, 2016; Hallinan, Frieclewald and McCarthy, 2012; Stoddart, Chan and Joly, 2016) , (2) technical data linkage issues (Kennedy and Millard, 2016), (3) Financial/budgetary obstacles to link administrative data to survey data.

Among the leading examples of successful linkage of administrative records and survey data, Sutherland et al. (2015) use data from the Millennium Cohort Study linked with administrative data from the National Pupil Database (sample of approximately 15000 students) to study the effects of socioeconomic deprivation in pupil achievement. Among the policy-relevant results, the study showed that Free School Meal eligibility remains a good measure of socioeconomic status in terms of explaining the levels of achievement in key stage 4.

This brief overview of the use of administrative data in single access research projects points us towards several opportunities and challenges that characterize the use of administrative data in social science research. The challenges include:

I. for the researchers, access and accreditation conditions that are often not clear;
II. transnational access – in most cases, researchers are still required to travel to get onsite access after the research project is approved. This leads to few studies conducting cross-national comparative analysis;
III. researchers may face a lengthy application period (which can require the attendance of training sessions);
IV. data privacy and protection: access to controlled/ secure access data (potentially disclosive data containing personal data) usually requires that the outputs are controlled – statistical disclosure checks – before being released from the secure environment.

On what concerns the opportunities:

i) new research possibilities derived from linked data (study of small social groups or minorities; new research topics; analysis of longer time periods);
ii) consolidation of best practices, adding scientific value to administrative data;
iii) technological innovations generate new possibilities in terms of data dissemination and linkage for both data providers/ controllers and researchers (e.g. secure virtual access).

Data Access Infrastructure

In response to the limitations identified with single access projects, several initiatives have been developed to provide improved access to administrative data for academic research. Unlike single access projects, these projects seek to develop reusable and scalable means for facilitating administrative data access. These are evident at both the national and international level but share a common focus on administrative data access as a service. The initiatives covered here can therefore be seen as the development of research infrastructure for administrative data. Here we provide a summary of some of these initiatives.

The Administrative Data Research Network (UK)

One of the most prominent recent initiatives, based in the UK, is the Administrative Data Research Network (ADRN), which was created in 2013 and funded by the Economic and Social Research Council (ESRC). This Network was a "partnership between UK universities, government, national statistical authorities, funders and research centres"[11], created to provide access to de-identified linked administrative data to approved researchers.

On what concerns the process of accreditation, the researchers submitted a research proposal to the Network that was then evaluated by the Approvals Panel, constituted by independent experts. If the proposal was approved, the Network trained the researcher[12] in order to ensure that the researcher is aware of the importance of - and takes measures to guarantee - the protection of privacy and confidentiality of the data.

The ADRN negotiated the access and data linkage process with the relevant government bodies, making the data available to the researcher in one of the secure environments of the Administrative Data Research Centres (England, Northern Ireland, Scotland and Wales), in the facilities of the data custodian or in other approved secure facilities in the UK[13].

When the researcher concludes the analysis of the data, the ADRN staff checked the findings for any statistical disclosure issue[14]. After these checks, - and only when the document was considered safe - it was released to the researcher. In order to showcase the potential of social science research using administrative data, the ADRN organized an Annual Research Conference[15].

The research projects approved and supported by the ADRN covered diverse areas, from crime and justice, to health, housing, education, employment and well-being. An example of one of these studies with policy-relevant findings was conducted by McGrath-Lone and colleagues. These researchers studied the experiences of children in care, using

---

[11] See https://adrn.ac.uk/for-the-public/faq/?About-the-Network.

[12] More details about the 'Safe Users of Research data environment Training' available via https://adrn.ac.uk/understand-data/sure-training/.

[13] List of approved secure facilities in the UK: https://adrn.ac.uk/get-data/safe-centres/.

[14] To clarify, statistical disclosure control corresponds to a set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations (OECD, 2005).

[15] UK Administrative Data Research Network Annual Research Conference: http://www.adrn2017.net/.

administrative data from the Department of Education ('Children Looked After Return' data and 'National Pupil Database')[16]. Among the key findings, the researchers discovered that children of black ethnicity are five times more likely than other children to have spent time in care; also, one third of children leaving care re-enter within five years (McGrath-Lone et al., 2016). The findings have been used to help social workers and local authorities to understand which children are most vulnerable and likely to go back into care.

In addition, there is a study by Brewer and Cribb (2016) focusing on UK time-limited in-work credits and how effective these are in helping lone parents. The authors use administrative data from the UK Department for Work and Pensions (Work and Pensions Longitudinal Study) to analyse the impact of two different benefits to lone parents who had previously been on welfare: 'In-Work Credit' (IWC) and 'Employment, Retention and Advancement Demonstration' (ERA). The policy-relevant findings of this study show that these programs contribute to increase job retention rates and "to large falls in the proportion of eligible lone parents on welfare, and large rises in the proportion in full-time work" (2016: 38).

The ADRN was discontinued in 2018 due to a lack of progress and the low number of approved projects. There was considerable demand from researchers for use of the network, but administrative data providers were slow to clear data for use, resulting in a very low number of successful linkage projects. In contrast to many European countries, administrative data remains under the control of separate ministries and departments as there is no centralized registry in the UK. This leads to data silos which do not interact well and largely inaccessible for the research community. The organization of administrative data access in the UK is therefore undergoing extensive reorganization to try and address these issues. At the time of writing, details of how this will be managed in the future were not available[17].

Le Centre d'Accès Sécurisé aux Données

In France, there is a remote access system created to facilitate access to confidential microdata - CASD, Le Centre d'Accès Sécurisé aux Données. Frequently, researchers need to have access to two or more datasets in their research projects – these data might also have been produced by different organisations and have different access requirements. So far, there have been little advances in terms of initiatives dedicated at facilitating administrative data matching in the social sciences context.

However, in October 2016, a new legislative framework was implemented in France in order to enable data matching with a specific code derived from the NIR (the national identifier) – a 'hashed NIR'. Given the need to protect confidential personal data, there is no inverse procedure that allows the initial data to be retrieved from the hash-code. When researchers request access to linked confidential datasets, the data custodian adds the hashed NIR and removes the NIR from the file. Then, the file without the NIR is transmitted to a trusted third party which is in charge of matching the files and making the merged data file available via a safe data centre.

[16] For more detail see: https://adrn.ac.uk/research-impact/research/children-in-care/.

[17] https://esrc.ukri.org/research/our-research/administrative-data-research-uk/

There are several examples of studies benefiting from CASD, covering diverse themes from productivity to innovation, work relations to immigration (to name a few). For example, the Observatoire Français des Conjonctures Économiques (OFCE) studied the between and within firm pay inequality with administrative data made available by CASD. Another research project developed by the Institut National d'Études Démographiques (INED) focused in the study of the characteristics of the immigrant population with Turkish origin, while a project by Université Paris 7 focused on gentrification and ageing in rural France.

ODISSEI

The UK is relatively unique in organizing administrative data access via its scientific funding council. In most European countries, administrative data access is coordinated via the statistical office as the data controller for government data. In the UK, the Office for National Statistics does not provide this function and therefore has a less prominent role in access provision.

In the Netherlands, CBS (the national statistical institute) produce a detailed catalogue of microdata[18]. In terms of access policy, the CBS requires that the primary objective of the researcher's institution is conducting statistical or scientific research and the results of the empirical analysis must be made public. The researcher can only access the files necessary for the approved research project in a secure virtual environment. If the researcher wants to download any output from the empirical analysis conducted in the secure environment, then a request should be made to CBS. The output is then checked by CBS staff and released to the researcher if there are no statistical disclosure issues. This Microdata Access Service Program currently facilitates around 500 projects per year in a diverse range of topics.

To further advance use of administrative data in the Netherlands, the Social Science Community established a national infrastructure. The Open Data Infrastructure for Social Science and Economic Innovations (ODISSEI) is responsible for enhancing and developing the use of linked administrative data in the Netherlands. It does this in four ways. Firstly, ODISSEI provides subsidized access to Microdata for institutional members of ODISSEI. Secondly, ODISSEI operates a series of open calls in which researchers can apply for free access to the administrative data, including the linkage of Survey data. Successful projects are then provided with full data support and stewardship throughout their project at no cost. These grants are particularly targeted at Early Career Researchers who find such access hard to achieve and resource.

In addition to this, ODISSEI is responsible for coordinating social science data collections in the Netherlands and actively promotes the linkage of survey data such as SHARE, ESS and the GGP to the administrative records held at CBS. Finally, ODISSEI has also been responsible for ensuring that the administrative data held at CBS can be analyzed within a secure High-Performance Computing Facility. Given that the administrative data held by CBS is detailed and complex, it can be exceptionally computationally demanding to analyze. To mitigate computational constraints, ODISSEI has created the ODISSEI Secure

[18] See https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/catalogus-microdata.

SuperComputer which allows for linked CBS data to be analyzed in a secure remote access environment on the Dutch National Supercomputer at SURFsara called Cartesius[19].

The ODISSEI Secure Supercomputer can be used to analyse highly sensitive data in a secure environment, including administrative microdata on persons, households, companies etc. from CBS. This is due to the system's architecture: SURFsara acts as Trusted Third Party between the data provider and the researcher, and analysis environment is strictly controlled and shielded. The ODISSEI Secure Supercomputer facilitates analysis of any sensitive data, and research communities ranging from humanities to health sciences have shown interest. Three pilot research projects have been successfully completed thus far and ODISSEI provided the following descriptions below.

Pilot Project 1: New Dimensions in Geospatial Analysis

The interest in neighbourhood effects goes back decades, but there is a lot of discussion on their importance and magnitude. There are two problems in the literature. First, neighbourhoods are often defined by using standard administrative neighbourhoods, such as census tracts, postal codes or other administrative units. The size of these units can vary from several hundreds of people to tens of thousands. Second, the neighbourhood context is often measured at one point in time, overlooking that the spatial context affects people over the course of their lives. This ODISSEI pilot project from TU Delft delivers both a conceptual and methodological contribution to the literature by developing a multi-scale and longitudinal approach to measuring bespoke spatial contexts of people. A multi-scale approach allows researchers to assess which spatial scale is more important for which people, and at which point in their lives, and whether there are differences between cities and urban structures in these spatial context effects. Analyses of heterogeneity in intergenerational mobility across spatial contexts or in very specific subpopulations become possible on a scale that is unprecedented in social research.

The researchers used the individual-level characteristics of every person in the Netherlands. These grid cells were then used to construct spatial context characteristics (for example on poverty) for 100 different spatial scales surrounding every individual. Such analyses have been carried out before, but often using highly selective or problematic data sources, and rarely covering a whole country. Statistics Netherlands microdata allow much more precise and comprehensive analyses. Starting from very small 100 by 100 meter grid cells, a distance profile of the residential context was calculated for each individual up to 10 by 10 kilometres. For a single indicator, this process would take months to calculate in the standard Statistics Netherlands' secure micro-data facility. There are nearly 400,000 inhabited grids cells in the Netherlands and the calculation of longitudinal entropy values of distance profiles generates more than 700 million new data points. The parallelisation that is made possible in the ODISSEI Secure Supercomputer reduces this process from months to hours, allowing much more experimentation with different operationalisations, and additional insights to be published in less time.

[19] https://userinfo.surfsara.nl/systems/cartesius

Pilot Project 2: Insights from Combining Socio-genomics and Administrative Data

Reaching out to disciplines such as behavioural genetics, socio-genomics, and genetic epidemiology is of in-creasing importance for the social sciences. However, linking genome-wide genetic data, or transcriptomic, epigenetic or other 'omics' data to social science data poses a set of challenges: the sheer size of the genetic data, computing requirements, privacy concerns and the requirement that data available at Statistics Nether-lands could thus far only be analyzed within its own secure remote access facility. In this ODISSEI pilot project, the Netherlands Twin Register (NTR), Statistics Netherlands and SURFsara developed algorithms and solutions for these challenges. NTR collects genetic data that currently contains over 40 million data points ('Single Nucleotide Polymorphisms', SNPs) per individual. These data had to be linked with data collected at Statistics Netherlands.

Linkage of phenotype and register data was done using the standard pseudo-anonymzation of person identifiers at Statistics Netherlands (see section 2.4.9.3) for participants in the NTR who had provided their consent. This procedure was first tested by the NTR researchers in a proof-of-concept study for NTR data on genetic liability to schizophrenia and Statistics Netherlands data on population density (Colodro-Conde et al., 2018), resolving a longstanding question in psychiatry about the aetiology of the relation between schizophrenia and urbanicity. Next, a study with full genome-wide data was done. An additional privacy protection challenge for working with these data is that they could be not organized through classical linkage keys. These data need to be analyzed with a defined order of subjects in the data file, thereby potentially creating a means to identify persons. To solve this problem, a pseudorandom record shuffling procedure was developed. This was successfully executed by SURFsara as a Trusted Third Party in such a way that neither the NTR researchers nor those involved at Statistics Netherlands could relate the identities of the subjects to the data.

The NTR pilot study was tailored to the current possibilities of the ODISSEI Secure Supercomputer and en-tailed a Genome-Wide Association study (GWAS) including millions of association tests between the genetic data (SNPs) and health care expenditures held at Statistics Netherlands. The sample with genotype data that could be linked to Statistics Netherlands data consisted of ~15,000 individuals and the GWAS was an important step in successfully demonstrating that linking genetic data to administrative data is feasible. This opens up an unprecedented range of new research areas as many longitudinal panels in the social sciences have now begun to include genotype information (as well as other big data such as Magnetic Resonance Imaging (MRI) and recordings from wearables). In this proposal, the ODISSEI Secure Supercomputer will accommodate these data and make it possible to link multiple longitudinal panels simultaneously to integral information at Statistics Netherlands. Running new and innovative genetic analyses on multiple traits, outcomes over time and gene-environment interactions will provide the opportunity to obtain insights into the role of genetics in social phenomena within Dutch society. For example, it will allow researchers to track genetic effects related to mental health on social inequalities in educational and health outcomes throughout the lifespan and allow the investigation previously established associations that might reflect gene-environment interplay rather than causal effects.

Pilot Project 3: Whole Population Network Analysis

One of the potentially powerful applications of administrative microdata is to examine how people are linked together through their work, education, neighbourhoods and families and how these networks structure re-sources and inequalities. This has normally been done on a limited scale, for example looking at how parents and children are correlated on specific variables or assessing the difference in outcomes between siblings. The integration of Statistics Netherlands' whole population data into the ODISSEI Secure Supercomputer has re-moved this limitation. Researchers from Statistics Netherlands are now able to analyse individuals' whole network of connections. The researchers have created a dataset comprising all 17 million residents of the Netherlands and have linked every person to their neighbours, their colleagues, their relatives and people whose children attended the same school. This is made possible by the high quality, longitudinally consistent identifiers that capture not only individuals but also the institutions to which they are connected. The research team created a dataset of over 800 million links which describe the latent structure of social network opportunities within the Netherlands.

The created network opens-up numerous research possibilities: e.g. on segregation/integration of networks of different social groups, inequalities, but also on vulnerabilities in network opportunities, on 'contagion' of social behaviours (such as fertility, divorce or crime), and so on. As a first example of the possibilities, the researchers investigated exposure to others with a different migration background. They applied a random walk procedure over the network for every individual in the country. In effect, this procedure would select a connection within someone's network at random and then see how many connections would have to be 'travelled through' until someone from a different migration background is encountered. This random walk procedure can summa-rise how 'exposed' an individual person is to persons of a variety of different backgrounds. Random walk analyses are not unusual in network analysis but were never performed on this scale: the team was able to investigate for the entire population of the Netherlands how this exposure to various groups varied between age groups, genders, or even specific towns and cities.

Data Without Borders

In addition to national level initiatives there have been several international iniatives to facilitate administrative data access. 'Data without Borders' (DwB) was created in 2011. This project came to a formal end in April 2015. One of the main objectives of this initiative was to facilitate the use of official microdata for the European Research Area. To this end, DwB:

> "worked towards preparing a comprehensive European service with better and friendly metadata, a more harmonized transnational accreditation and a secure infrastructure that would allow transnational access to the highly detailed and confidential microdata, both national and European, so that the European Union would be able to continuously produce cutting-edge research and reliable policy evaluations" (DwB webpage).

With these priorities in mind, the network had to be able to manage different legal frameworks across Europe. Among the main contributions of DwB, Silberman et al. (2015) emphasise usability (the virtual research environment should provide every standard tool for processing microdata), collaboration (users should be able to share data and results according to certain rules – legal and organisational) and management tools (Research Data Centre (RDC) staff should be able to manage the network).

The DwB established a broad partnership of three communities, including 29 partners belonging to the European Statistical System (national statistical institutes or statistical departments); Council of European Social Science Data Archives (CESSDA); and the research community (universities).

In the period 2012-2015, the DwB invited academic researchers resident within the European Union member states or European Free Trade Association associated countries to apply for access to highly detailed microdata from RDCs in France, Germany, Netherlands and the UK (researchers applied to access specific datasets at one or more RDCs not situated in their country of residence). During the eight calls, the expert User Selection Panel approved 40 research projects.

Some RDCs provided remote access to the data in secure virtual environments, yet some other researchers had to travel to receive training and conduct the data analysis onsite. An example of a publication deriving from a DwB approved project is the study by Bach et al. (2015) on the influence of different information sources on innovation performance. The study used detailed CIS (Community Innovation Survey) data.

The DwB team also evaluated the state of the informational infrastructure and in particular remote access systems and produced a report on the 'State of the Art of current Safe Centres in Europe'. In terms of user access, one of the conclusions is that remote access to controlled data has proved to be among the best alternatives to provide secure access to researchers (Gurke et al., 2012). On the other hand, the traditional solution has been onsite access, which requires that both national and foreign researchers travel to safe data centres to analyse the confidential microdata. The report also mapped out remote access providers and their strategy and practices (eight providers - statistical institutes and data archives- in seven European countries):

1. Statistics Denmark
2. Office for National Statistics (UK)
3. UK Data Archive – Secure Data Service (UK)
4. Statistics Sweden

5. Statistics Slovenia
6. Statistics Netherlands
7. Secure remote access centre (CASD) (France)
8. Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research (Germany)

Alongside DwB, the ESSnet project 'Decentralised and Remote Access to Confidential Data in the European Statistical System (DARA) also focused on the goal of facilitating access to confidential microdata, exploring the feasibility of setting up a network of Safe Centres for research purposes in the European Union. One of the central recommendations deriving from this project was "to establish a network of Safe Centres, which are connected via remote access to a central node (Eurostat), where the data are stored on a secure server" (Brandt and Schiller, 2013: 2).

In terms of access services, the DwB produced a 'Database on National Accreditation and Data Access Conditions' that provided details on the access regulations to data produced by the leading statistical agency on each country (including modes of access, timing, application forms and other documentation). This database is now integrated in CIMES – Centralising and Integrating Metadata from European Statistics – and the objective of this initiative is to centralise relevant information that would otherwise be scattered across Europe, facilitating data and resource discovery.

Another contribution of DwB is the Microdata Information System for Official Statistics - MISSY, created to offer systematically structured metadata for official statistics. We should also note that the DwB produced the Legal Frameworks Visualisation Tool to help researchers explore the legal gateways to transnational and national data access.

DwB produced other products concerning data access to controlled official microdata:

i. Researchers' Needs regarding Secure Access to Official Microdata

The report on 'Researchers' Needs regarding Secure Access to Official Microdata' resulted from individual and group interviews of a sample of researchers with experience in secure data analysis (mostly in DwB RDC partners). The objective of this report was to assess not only researchers' needs, but also reflect on a possible architecture for an European Remote Access Network that takes into account the security demands specific to controlled data.

Among the needs identified in the report, researchers argue for an access system that does not require travel and that allows researchers to use their own equipment to access the data. An additional point raised concerned the merging of datasets and included a recommendation for a single point of access for the merging process. Finally, researchers emphasise that it is important that there are more homogenous practices in terms of the software used; improvements can also be made in terms of data documentation, support and a shorter turnaround time for output checking ; researchers would also like to be able to discuss their outputs with other researchers.

ii. Research Data Centres and ISO 27001 – A Guide

Regarding the DwB objective of enhancing information security, the report on 'Research Data Centres and ISO 27001' provided an implementation guide of the ISO 27001 standard. This document is particularly relevant to national statistical institutes and data archives, as it defines the guidelines to be adopted by data providers in terms of risk assessment and risk

management, as well as data and user controls. The report also presents the case of the UK Data Archive and the process to achieve ISO 27001 certification.

iii. Guidelines for Output Checking

The DwB 'Guidelines for Output Checking' intended to provide practical guidelines for output checking – which consists in the process of checking the disclosure risk of research results based on microdata files (Bond et al., 2015). The document includes rules for different types of outputs, along with recommendations and best practices for organisational and procedural aspects associated with output checking.

Among the activities and events prepared by the DwB team, there were a number of training courses and resources directed at the research community. Each three-day training course covered an individual dataset (EU-Statistics on Income and Living Conditions; European Labour Force Surveys; European Adult Education Survey; European Census Microdata) and also the legal requirements for data access. In order to stimulate further discussions between data producers, researchers and data providers, the DwB promoted two European Data Access Forum (in 2012 and 2015) and two User Conferences (in 2013 and 2015).

Facilitating International Data Access

Population Registers are of highest quality and most pervasive in Scandinavia and this is also where administrative data is most widely used in research. For example, Statistics Denmark, makie register data available to researchers affiliated with Danish research institutions[20]. Statistics Denmark allows researchers to apply for data linkage with other administrative data or their own data, after approval by the Danish Data Protection Agency. Denmark, Finland, Norway and Sweden have long-standing experience with register-based statistics, having introduced central population registers since the 1960s. The UNECE report (2007) 'Register-based statistics in the Nordic countries' compiled and reviewed some of the best practices with a focus on population and social statistics[21]. In 2015, these countries have joined efforts to create the Nordic Microdata Access Network (NordMAN). The national statistical offices agreed on a common model of cooperation, in order to integrate researchers' access to Nordic social microdata[22].

A network more orientated to non-register countries is the International Data Access Network (IDAN), which is a collaboration between administrative data providers in the UK, the Netherlands, France, and Germany[23]. The collaboration seeks to extend existing remote access facilities at institutions in each country to facilitate cross-national secure access. The project focuses on harmonizing access procedures and training across the four countries, but it is as of yet unclear as to whether the data will be analyzable in a single environment to facilitate comparative research.

---

[20] More details via http://www.dst.dk/en/TilSalg/Forskningsservice.

[21] According to the report, in 1981 "Denmark was the first country in the world to conduct a totally register-based census and Finland followed in 1990. From 1980, the censuses in Norway and Sweden have been partly register-based" (2007: 6).

[22] For more details on the access conditions see http://nordman.network/.

[23] https://idan.network

Summary

To conclude, in this section we have seen promising advances in terms of initiatives concentrated in facilitating researchers' access to controlled microdata, including efforts to extend the use of official microdata in the European Research Area (DwB) and a number of national level initiatives dedicated to enable research projects that require access to administrative data (namely, ODISSEI in the Netherlands; CASD in France; ADRN in the UK). However, these access provision initiatives still have a very limited capacity, in particular when considering research projects using cross-national data, reaching a relative small number of researchers: in terms of outcomes, only 40 research projects were approved by the Expert Panel and only nine publications by supported projects are listed in the DwB webpage[24]. DwB received funding from the EU 7th Framework Programme amounting to approximately five million euros, which gives an indication of the costs and challenges that are associated with setting up a network that supports empirical researchers accessing administrative data. On the other hand, the initiatives presented here are mainly designed focusing on a post-hoc data linkage, where the researchers have no possibility of interacting or familiarising themselves with the data when planning their research or applying for data access.

[24] http://www.dwbproject.org/access/outcomes.html.

Integration in social surveys processes

There is an increasing necessity to reflect on the potential benefits and challenges that additional data linking with other data sources could provide. We focus here on three social survey programmes – European Social Survey (ESS), Generations and Gender Programme (GGP) and the Survey of Health, Ageing and Retirement in Europe (SHARE) – that have been matching additional data to survey microdata in order to enrich the research data available, which represents the core foundation for many cutting-edge scientific research projects and policy reports.

The opportunities that this type of integration can allow have been diagnosed by many in recent years (Andersson et al, 2014; Card et al, 2010; Chetty et al, 2011; Connelly et al, 2016; Emery, 2016), in a context where these extended research possibilities are recognised, alongside with the challenges that low response rates in surveys and rising fieldwork costs represent. It is important to note also that the integration of administrative or other types of data with survey microdata can be designed to be completed before or after the survey process. The ESS and GGP examples, discussed below, exemplify an ex-ante integration, while the SHARE project exemplifies an ex-post process.

The Generations & Gender Programme

Among the social survey programmes, the GGS (Generations and Gender Survey) is a longitudinal, cross-national survey experienced in the successful implementation of some initiatives linking survey microdata and administrative records. The Swedish GGS – part of the International GGP[25] – was fielded in 2013 with a response rate of 54.7% and 9688 respondents. The response rate is in line with the GGP average response rate of 55.7% (Emery, 2016).

Benefiting from the fact that Sweden has a central population register, the full sample of the Swedish GGS was linked to a wide range of administrative records before the fieldwork process (carried out by Stockholm University in collaboration with Statistics Sweden) – participation in the survey was dependent on respondent's consent to record linkage. This experience shows that despite the fact that population registers and some administrative data have highly restricted access, surveys can provide a 'key' to consent to allow for linking with surveys and even other data sources. In fact, this strategy allowed also for administrative data validation "this basis of linkage consent enabled the fieldwork to pre-load administrative records […] enabling respondents the opportunity to correct the data where they deemed necessary" (Emery, 2016: 6) – as an example, 18.3% of the respondents corrected the educational level information recorded in the administrative records during the interview.

This initiative shows that, even in countries as Sweden with advanced population registers and social statistics, the incorporation of administrative data in the survey process (pre-load administrative records) benefited both data infrastructures and data users. In addition, there are also important substantive reasons to integrate administrative data in social survey processes as this linkage provides new research opportunities otherwise not available.

Furthermore, the Austrian GGS incorporated administrative data in the survey process with a view to analyse attrition (due to non-contact or due to non-cooperation) between the first and

[25] http://ggp-i.org/.

second wave[26]. Given that Austrian legislation requires that individuals notify the authorities about any residential move, the central register is continually updated. This way, if respondents moved between wave 1 and 2, the contact address in the register was updated (attrition due to unknown address was expected to be comparatively low) - it was crucial for panel maintenance that Statistics Austria had access to the central register. Buber-Ennser (2014) concluded that the Austrian GGS panel has a relatively low level of attrition (22%), with a small bias towards family-oriented persons, as well as respondents with lower educational levels and those with migration background.

European Social Survey

The ESS developed the ADDResponse (Auxiliary Data Driven non-Response bias analysis) project which uses different sources of geocoded auxiliary data and survey paradata to understand and correct for nonresponse bias in the UK wave of ESS (Butt and Lahtinen, 2016). The geographical identifiers and small-area administrative data included 2011 Census socio-demographics, crime rates, indices of deprivation, electricity consumption, school absences, benefit claimants, wellbeing, and voting in local elections. These auxiliary data was collected from three different sources – i) small-area administrative data, ii) commercial marketing data and iii) geo-coded information from the Ordnance Survey on the location of the sampled addresses. The project linked around 400 variables from different data sources to the ESS round 6 UK sample of addresses[27].

One of the main purposes of the project was to explore nonresponse bias in social surveys. This is considered one of the major challenges facing empirical research presently as nonresponse can produce significant bias into the conclusions drawn from survey data. However, in order to address this challenge, researchers need more information about respondents and non-respondents, but they also need to have data about the extent and sources of nonresponse bias. In face of declining survey response rates, the project team explored the causes and correlates of nonresponse. The ESS has previously done research on nonresponse (Stoop et al, 2010), but this has relied mostly on paradata or auxiliary information collected by the interviewer. In contrast, this project exemplifies ex-ante integration of these type of data in survey processes. The idea is that the lessons learned from this process will inform next rounds of fieldwork and improve data collection practices (potentially reducing non-response rates). In fact, the integration of different data sources in survey processes can be seen as a 'virtuous circle' – the survey process can help complement/ validate administrative and other type of data, but also the integration of administrative records with survey microdata is of interest because that way the data is enriched and supplemented and that will benefit both administrative data holders and social scientists.

---

[26] The first wave was carried out in 2008/9 and the second wave in 2012/13. In total, 3907 interviews with wave 1 respondents could be realized in wave 2 (panel stability of 78%).

[27] Further details about ADDResponse available via https://blogs.city.ac.uk/addresponse/. The data can be accessed via the UK Data Service (SN8066), through Secure Access conditions (requires accreditation as an ESRC Accredited Researcher, completion of face-to-face training, and agreement to the Secure Access User Agreement and the License Compliance Policy). A variable list for this database is available via http://doc.ukdataservice.ac.uk/doc/8066/mrdoc/excel/8066_addresponse_variable_listing.xlsx.

Matching survey data with geocoded auxiliary data has attracted growing interest in recent years. Butt and Lahtinen (2016) argue that this multi-level multi-source (ML-MS) approach can provide valuable insights both regarding the substantive analysis, but also at the methodological level. Nonetheless, there are also some important challenges that are not negligible in projects of this type. Firstly, for international survey programmes, as the ESS, there are frequently issues with the availability (or lack thereof) of comparable auxiliary data cross-nationally - the definition and measurement of auxiliary variables will likely vary. An additional challenge concerns the need for data protection considerations (and the associated potential for deductive disclosure) given the level of detail of the data derived by the linking of survey microdata with detailed auxiliary data (including low level geographical identifiers). In the case of the ADDResponse project, the linked data is deposited securely with the UK Data Service and the access to the dataset is restricted.

Lahtinen and Butt (2016) analysed the nonresponse behaviour in the ESS UK wave using both external auxiliary data and survey paradata. Interviewer observations were considered helpful in predicting nonresponse, while external auxiliary data was also regarded as relevant, but there were some quality issues with commercial variables (Lahtinen and Butt, 2016). This paper also showed that further analysis is required, eventually using different modelling techniques, as a robust and strongly predictive model of response propensity remains elusive (Lahtinen and Butt, 2016). Overall, the project team found that the auxiliary provided "little added value [and the team] failed to identify suitable POIs for nonresponse analysis" (Butt and Lahtinen, 2016a: 24).

The Survey of Health Ageing & Retirement in Europe

Another example is the work done by SHARE - an international panel study that focuses on health, socio-economic status and social and family networks of individuals aged 50 or older. Currently, SHARE covers 27 European countries and Israel and the data is made available to the research community free of charge[28].

The project linking SHARE data with administrative records of the German Pension Insurance scheme (DRV) started as a pilot study in 2009 and is now a standard module of the German SHARE (designated SHARE-RV) – this link to the administrative records is permanent (the data can be updated well after the fieldwork was conducted). The survey respondent was asked whether she/ he consented to the linkage of personal German Pension Insurance data to be linked to the SHARE interview. This is a direct linkage, meaning that the data of the same person (German SHARE respondents) were linked using the respondents' Social Security Number (SSN) as a unique identifier (SHARE, 2017). SHARE-RV allows the investigation of a number of aspects related to the respondents' working history and their socio-economic status at a later stage of their lives, enabling also the analysis of the circumstances of other persons living in the same household (linkage of administrative records)[29].

As an illustration of the potential of the SHARE data linked with these administrative records, Börsch-Supan et al. (2015) used the data to evaluate a reform introduced by the German

[28] Further details about SHARE: http://www.share-project.org/home0.html.

[29] See http://www.share-project.org/special-data-sets/record-linkage-project/record-linkage-share-rv.html.

government. This policy re-introduced early retirement at age 63 for workers with 45 years of contributions to the pension system and help underprivileged workers who had longer working histories (typically in less well-paid and more physically demanding jobs). These two types of data "combine the best of both worlds: administrative data have very precise information on employment history and resulting pension claims while SHARE offers data on socio-demographics which are not available in administrative data" (2015: 276).

The authors analyse whether this reform achieved its goals, using data on health and socio-economic status of the eligible workers to see if in fact this policy was helping those in more vulnerable circumstances. The findings seem to point in a contrary direction: if the objective of the reform was to help those more deprived and with poor health, then there seems to be no evidence that the policy achieved this goal.

In contrast, Bingley and Martinello (2014) used SHARE data linked to Danish administrative registers to assess measurement error in SHARE. This validation study compared the representativeness of SHARE data in terms of education level, income and employment. In this research, the authors were able to link 1670 out of 1707 respondents (98% of the sample) from the first wave of SHARE in Denmark to the register data. For this study, the authors had to present a research project and seek permission to have access to the administrative data (in this case, a remote desktop access application). Furthermore, in order to be able to access the data, they had to sign a confidentiality agreement, install a VPN client and receive a security token look-up before gaining permission to remotely access the data.

This study can be classified as an internal validation study, as it takes a sample of a survey and compares it to an official validation source – which is assumed to be measured without error (Bingley and Martinello, 2014). The findings reveal that measurement error in schooling, labour market status and income is modest and insignificant. The study found that SHARE respondents with higher reported or registered incomes tend to overstate their level of education. On average, gross household income is not statistically different in SHARE and register records. In general, labour market status is reported in line with the information in the register (misclassification probability decreases with age, i.e. younger respondents are more likely to respond at odds with registers) (Bingley and Martinello, 2014).

Summary

To conclude, the integration of administrative data (and other types of data) in social surveys programmes proves that there are important gains in these procedures, from a substantive point of view, but also from a methodological perspective (e.g. Austrian GGS and ESS ADDResponse project). Social survey programmes and data owners need to take these opportunities seriously as the expansion and improvements in data availability are crucial for not only for decision-making and evidence-based policy evaluations, but also for path breaking social science research.

Infrastructures with statistical disclosure controls

National statistical agencies and other data providers have as one of their core missions to provide direct and secure access to a wide range of data resources, some of which can be very detailed and contain identifiable data. In general, these data are located in a secure server and involve stringent conditions of access.

Given this context, data owners and providers need to develop strategies and implement a number of measures to protect the confidentiality of the data subjects, including statistical disclosure control (SDC) aimed at minimising the risk of disclosing information on individuals, businesses or other organisations (OECD, 2005). After a successful application process, researchers usually need to attend training sessions and any output derived from their statistical analysis is screened using SDC methods by the data provider staff. When the output is considered safe, the results are released to the researcher.

As Hundepool et al. (2012) note, there are three main reasons why confidentiality protection is crucial:

i) a fundamental principle for official statistics is that the record of individual persons, businesses or events used to produce official statistics are strictly confidential and should be used only for statistical purposes (in accordance with the 1992 UN Economic Commission report 'Fundamental Principles for Official Statistics' and the European Statistics Code of Practice);
ii) in many countries, there may be a legal obligation on a national statistical institute to protect personal data and individual businesses[30];
iii) data security and trust by survey respondents: "one of the reasons why the data collected by national statistical institutes is of such high quality is that data suppliers or respondents have confidence and trust in the NSI [national statistical institute] to preserve the confidentiality of individual information" (2012: 8).

The IPUMS-International (Integrated Public Use Microdata Series) project[31] - led by the University of Minnesota Population Centre and launched in 1992 - is a leading example of the efforts necessary to create a global resource for restricted access census microdata. Traditionally, census data have been available only in aggregated tabular form, which limits the research potential of the data.

The contribution made by IPUMS-International is very important because the project has collected the world's largest archive of publicly available census samples[32], negotiating agreements with national statistical offices to disseminate integrated census microdata to accredited researchers. The project team works in data integration, analysing the structure of the data, applying internal consistency checks and correcting any errors (Ruggles et al., 2015). IPUMS makes data available free of cost to researchers who apply with a viable research project. The data disseminated covers a broad range of population characteristics, including household composition, ethnicity, migration, fertility, nuptiality, education, occupational structure, among other.

[30] Hundepool et al. (2012: 7) also specify that "where public statements are made about the protection of confidentiality or pledges are made to respondents of business or social surveys these place a duty of confidence on the national statistical institute".
[31] More details on this project available via https://international.ipums.org/international/about.shtml.
[32] In 2011, the data infrastructure included 397 million anonymized, integrated person records representing 100 million households in 62 countries and 185 censuses (Esteve, 2011).

On what concerns the access policy, IPUMS-International has a number of technical, legal and administrative controls to assess users' applications with the ultimate goal of protecting the privacy and confidentiality of the microdata. Researchers need to fill in an application form detailing the intended usage and must agree with specific conditions of use of the microdata. Each researcher specifies the required data, indicating the country (or countries), census year(s), variables, sub-populations and sample density[33]. This way, no two data extracts are alike and this method of dissemination has already "weathered the test of time, and indeed as usage soars, the rapid acceleration of internet transmission speeds has validated the IPUMS approach." (Esteve, 2011: 3).

In terms of disclosure control measures, besides the detailed, legally binding electronic application, researchers must also agree to abide by the restrictions defined in the conditions of use[34]. There are also statistical control measures, which involve the suppression of records by subsampling; the suppression of names and geographical details (place of birth or residence); some variables may be top-coded or globally recoded, deletion of digits for some variables (e.g. occupation, geography) or the suppression of a variable/ variables. An additional statistical control measure is provided by randomly ordering the records and swapping the geographical identifiers of an undisclosed number of households (McCaa and Esteve, 2012).

In terms of technical measures to ensure the secure transmission of the microdata, after the preparation of the data extract, the approved researcher is prompted to retrieve the data from a password-protected page within 72 hours – the data are encrypted during the transmission using 128-bit SSL standard[35].

In terms of limitations of this project, the fact that each data extract is unique can also be seen as a potential limitation as this prevents the replication of the empirical analysis conducted. A further limitation derives from the fact that, depending on the date of the last census, the data available for a given country can be almost ten years old – it should be noted that this limitation derives more from the timing of census data collection than from the project itself.

Furthermore, the role that researchers play in safeguarding the confidentiality of highly detailed controlled microdata should also be taken into account. In fact, researchers as active data users can play a very important role in terms of promoting responsible data usage and protecting data confidentiality - this is certainly in their long-term interest. In fact, several social scientists have advocated that data access is a social good, defending that the rush to ensure complete levels of privacy in the research context paradoxically results in less social benefit, rather than more (Madsen, 2003). Madsen (2003) identifies this 'privacy

[33] IPUMS microdata extracts allow for international comparisons as both microdata and metadata are integrated for all censuses and countries.

[34] These restrictions include: prohibition of redistribution, restriction to scholarly use, prohibition of commercial use, protection of confidentiality, assuring security, enforcing strict rules of confidentiality, scholarly publication permitted, correct citation of the microdata, threatening disciplinary action for violations and reporting of any errors discovered.

[35] This model contrasts with other infrastructures that do not permit any download from the remote access system (e.g. UK Data Service Secure Lab).

paradox' and argues that it derives from a narrow conception of the right to privacy and from an uncertainty regarding data ownership (data constituting private or public property).

Ruggles et al. (2015) compiled the usage statistics of IPUMS and the figures are impressive: 10,000 researchers registered to use the international IPUMS data, producing approximately 1,000 articles and working papers. The authors note also that the largest disciplinary group of users does research on Economics.

To exemplify the diversity that characterizes the extensive usage of IPUMS microdata, we have selected four studies concentrating on different thematic and geographical areas. Esteve et al. (2012) investigate the household formation patterns of young women (25-29 years old) in 13 Latin American countries since the late 1960s using census extracts from IPUMS. The article advances a new typology for classifying individuals in various household living arrangements, finding a mixed picture among the 13 countries "with countries such as Argentina, Brazil, Chile, Costa Rica and Uruguay following a more European pattern and the Central and Northern Andean countries maintaining stronger traditional family cohesion" (2012: 724).

It is important to reflect on the importance of census data for social science research. These data are collected by national government agencies, typically every 5 or 10 years, and consist of a complete enumeration of the population, with the objective of describing the socio-economic and demographic profile of the country. Among other purposes, census statistics provide crucial information that is used by government departments to allocate funding to public services. The access to the census data is usually restricted. However, it should be noted that there have been important efforts to facilitate access to census data, in particular the initiative by IPUMS-International (Integrated Public Use Microdata Series)[36] – and, in the UK, the UK Data Service Census Support[37].

As an example of a scientific study using this type of data, Clara Mulder's ERC-funded project ('Family Ties') studying internal migrations and its labour market outcomes, identifying the role of family ties in internal migration, immobility and labour market outcomes. Previously, Schaake, Burgers and Mulder (2013) used Housing Research Netherlands data (from the Dutch Ministry of Housing, Spatial Planning and the Environment) to study whether ethnic differences in the change in socio-economic status and ethnic composition of movers' neighborhoods can be attributed to respondents' education and income.

---

[36] IPUMS International data is restricted to scholarly and educational purposes and is "the world's largest collection of publicly available individual-level census data. The data are samples from population censuses from around the world taken since 1960. Names and other identifying information have been removed. The variables have been given consistent codes and have been documented to enable cross-national and cross-temporal comparisons". (https://international.ipums.org/international/overview.shtml). In point 4.3, we will discuss in detail the access policy and statistical disclosure controls put in practice in this project.

[37] See https://census.ukdataservice.ac.uk/. This service provides open access to aggregate UK census data (InFuse and Casweb), while individual and household files can only be accessed in safe settings by approved researchers.

Also, Anderson et al. (2014) used German and Danish register data to study the role of female labour-market attachment and earnings in childbearing progressions. Using event-history techniques, the authors tested if the welfare state shapes the female earnings and fertility nexus. They conclude that “in countries like (West) Germany, where work and family life have been rather incompatible, female earnings should be negatively associated with having children. In countries like Denmark that support maternal employment, women will be more inclined to have children once they have established themselves in the labour market” (2014: 310).

The study by Eichenlaub et al. (2010) uses U.S. census data to study the impacts of the Great Migration (migration flux from the south of the U.S. beginning in the 1910s and ending in the early 1970s), comparing migrants who left the south with their southern contemporaries who stayed behind, both those who moved within the South and the sedentary population. The findings challenge the widespread view regarding the benefits of leaving the south, “recognizing the complexity of the migration decision and offers a fuller appreciation of causal influences that transcend the individual migrant” (2010: 120).

Another example of a study using IPUMS census data is Zueras and Gamundi (2013), who investigate the expansion of the number of persons aged 65 to 84 years old living alone in seven European countries (Spain, France, Greece, Hungary, Portugal, Romania and Switzerland). According to the authors of the study, the proportion of elderly living alone increased in 2001 in all countries but Romania. Among the main drivers of this phenomena, it is argued that socio-economic variables play a particularly important role (and, more specifically, the educational level of the individuals).

Finally, Lam and Marteleto (2008) focus on the three stages of demographic transition from a child’s perspective, each with different implications for resource competition at the family and population level. The authors use IPUMS census data to analyse changes in fertility, mortality, surviving family size and cohort size in eight countries (Brazil, Costa Rica, Ecuador, Kenya, Mexico, South Africa, Uganda and Vietnam). The researchers note that “during the two or three decades of the demographic transition that most countries spend in Stage 2 [the stage in which children compete for resources with fewer siblings], children benefit from reduced competition for resources inside the family but face increased competition at the population level” (2008: 249). Lam and Marteleto (2008) argue that the implications of the changes observed in family size and cohort size have important impacts in a number of outcomes such as schooling, health and even labour market experience.

Future Technical Developments

In addition to statistical disclosure controls, there are a number of technical developments which could circumvent the legal and ethical issues associated with linking and analyzing administrative data. One of these concerns the use of distributed computation through federated learning algorithms. This has been applied in the Personal Health Train which is a concept that was developed in the medical sciences. In the Verantwoordelijke Waardecreatie met Big Data project, funded by the Dutch National Science Agenda (project number: 400.17.605) Maastricht University together with De Maastricht Study and Statistics Netherlands are implementing the Personal Health Train on health data. Its basic concept is that several stakeholders want to collaborate in data analysis, but do not want or are legally not allowed to share the data with one another – often due to privacy considerations. Rather than appointing a Trusted Third Party, the researchers develop an algorithm (in this analogy: the train) that visits each stakeholder (‘station’), analyses the data on site and goes to the

next station with the analysis results, but without the data. The concept of distributed computation has been widely studied in the medical and computer sciences (Sun et al., 2018). Though having great potential in the social sciences, to date, no research has been done about its applications in this field.

It is important to consider the potential shifts in administrative data usage that could be opened by such technical developments. One similar example that is already in operation is the Coleridge Initiative in the United States[38]. This is an infrastructure which provides data owners within the computational infrastructure and expertise necessary to conduct analysis of administrative data whilst also providing a framework for data linkage and integration. The application of secure distributed computing to the Social Sciences in Europe is severely underdeveloped in this regard and there is significant potential that is not being utilized.

[38] https://coleridgeinitiative.org/

3 Legal and ethical challenges related to use and re-use of administrative data

Having provided an overview of existing administrative data linkage initiatives and examples of research, we now turn to the legal and ethical issues associated with administrative data linkage and the implications for the diverse forms of linkage. We structure this overview around the General Data Protection Regulation and the extent to which it now shapes administrative data linkage.

Informed consent issues

Article 4(11) of the GDPR defines consent as: "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her". In addition to this definition, several new requirements on consent follow from other provisions in the GDPR. In particular, article 7 sets out conditions for e.g. keeping records of consent, making requests for consent clear and concise, and the right to withdraw consent (ICO guidelines). When processing is based on consent, people will also generally have stronger rights, like the right to erasure and the right to data portability. If consent is obtained in full compliance with the GDPR, it should function as a tool that gives data subjects control over whether or not personal data concerning them will be processed (WP29, p. 4).

Legal grounds other than consent

In survey research, the respondent's informed consent usually constitutes the ethical and legal basis for the processing of personal data. Obtaining consent from respondents would also in many cases be a necessary prerequisite for surveys which directly link survey and individual-level administrative records. However, it could be difficult to manoeuver which and how much information to provide to the data subjects. Moreover, some situations make consent not feasible to obtain.

One scenario is admin data collected e.g. from population registers to form sampling frames for a survey. The purpose of this handling of personal data would for instance be to draw a representative sample, to recruit participants, and to get an overview of the total population. In such case, the handling of personal admin data happens *before* informed consents are collected from the data subjects. Thus, since consent is not yet appropriate as a legal basis, another lawful ground must be considered for this processing. Article 6 of the GDPR lists all together six lawful bases, whereas 6(1)(e) states that *"processing shall be lawful if necessary for the performance of a task carried out in the public interest".* Furthermore, Article 6(2) and (3) allow Member States to introduce more specific provisions, and state that the basis for public interest shall be laid down by Union or Member State law. The latter provisions may actually lead to various national implementations of the GDPR regarding whether (and to which extent) it is established that research purposes falls under the definition of "public interest".

An important notice is however that even if "public interest" is the appropriate lawful ground for the handling of admin data for sampling purposes, the other requirements of the GDPR still applies, e.g. regarding the rights of the data subjects. Any possible derogation from these rights must have a legal basis.

When the sampling is done, a typical next step is to invite the sample to participate in the survey. At this stage, information about which personal admin data have already been handled should be provided to the data subjects (Article 14(3)(b)). Thus, the further handling of admin data will be based either on consent (from the participants) or, for the non-respondents, continue to be grounded in "public interest".

Broad consent

Another scenario is when survey data are collected and stored with the intention of keeping the possibility open to future linkage to admin data for research purposes. At the starting point of such a project, it could be challenging to provide the respondents with all relevant information. One issue is how specific the information about future linking and research purposes must be to secure an informed consent. Another issue is to what extent future scientifically interesting admin data can be sufficiently covered. If admin records must be named to ensure an informed consent, this would limit future research if new, relevant admin data sources are established.

In essence, the GDPR puts great emphasis both in individuals having clear granular choices upfront and furthermore; ongoing control over their consent. Article 6(1)(a) confirms that a consent must be given in relation to "one or more specific" purposes and that a data subject has a choice in relation to each of them. This requirement aims to ensure user control and transparency for the data subject. As a part of "specific", the data subject must be given granular options to consent separately to different types of processing wherever this is appropriate.

The WP29 guidelines however comment that recital 33 "seems to bring some flexibility to the degree of specification and granularity of consent in the context of scientific research" (WP29, p. 27). Recital 33 states:

> "*It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose*".

The WP29 guidelines points out that research projects based on consent anyhow must have a well-described purpose, although it could be described at a more general level. Furthermore, the WP29 guidelines seem to clearly indicate that other controllers can rely on the original consent if previously named: "If the data is to be transferred to or processed by other controllers who wish to rely on the original consent, these organisations should all be named [in the process of obtaining consent]" (WP29, p. 13-14). How should naming of other controllers be handled if survey data are collected with the aim of making linkage to admin data available for the whole research society?

As stated by The Information Commissioner's Office and Finnish Social Science Data Archive (FSD), more detailed guidance on the naming of parties and the future purposes of the processing at the time of data collection is needed.

Transparency is an additional safeguard when the circumstances of the research do not allow for a specific consent. A lack of purpose specification may be offset by information on the development of the purpose being provided regularly by controllers as the research

project progresses so that, over time, the consent will be as specific as possible. When doing so, the data subject has at least a basic understanding of the state of play, allowing him/her to assess whether or not to use, for example, the right to withdraw consent pursuant to Article 7(3) (WP29, p. 28).

Information to the data subjects

Based on Article 5 of the GDPR, the requirement for transparency is one of the fundamental principles. Providing information to data subjects prior to obtaining their consent is essential in order to enable them to make informed decisions, understand what they are agreeing to, and for example exercise their right to withdraw their consent (WP29, p. 13).

Any information addressed to the data subjects and/or request for consent must be provided in a concise, intelligible and easily accessible form, using clear and plain language (Article (7)(2), 12 and Recital 78). For consent to be informed, at least the following information is required (WP29, p. 13):

(i) the controller's identity,

(ii) the purpose of each of the processing operations for which consent is sought,

(iii) what (type of) data will be collected and used,

(iv) the existence of the right to withdraw consent,

(v) information about the use of the data for decisions based solely on automated processing, including profiling (cf. Article 22)

(vi) if the consent relates to transfers, about the possible risks of data transfers to third countries in the absence of an adequacy decision and appropriate safeguards (Article 49 (1a)).

In addition to this list of required information for consent to be valid, a controller must also deal with the separate information duties laid down in Articles 13 and 14 in order to be compliant with the GDPR. In practice, compliance with the information duties and compliance with the requirement of informed consent may lead to an integrated approach in many cases. However, the WP29 guidelines points out that valid informed consent can exist, even if not all elements of Articles 13 and/or 14 are mentioned in the process of obtaining consent. These points should then be mentioned other places, such as in the privacy notice of a company. WP29 has issued separate guidelines on the requirement of transparency (WP29, p. 15).

Article 13 and 14 outlines the following list of required information:

Article 13 - Information to be provided where data is collected from the data subject

Data subjects (respondents/non-respondents) must be provided with the following information at the time data are obtained and when information is updated:

The identity and the contact details of the controller and, where applicable, of the controller's representative;

- Processors do not need to be named as part of the consent requirements, although a controller will need to provide (e.g. on its website) a full list of recipients or categories of recipients including processors (WP29 p. 14).

The contact details of any data protection officer;

The purposes of the processing for which the personal data are intended as well as the legal basis for the processing;

- Description of the purposes of the survey.
- Relevant legal bases for a survey infrastructure can be:
    - Consent (Article 6.1 (a)
    - Explicit consent (Article 9.2 (j) when collecting sensitive data, e.g.: “the questionnaire involves questions on e.g. political opinions, religious beliefs, health issues etc.”
    - Processing is necessary for a task carried out in the public interest (Article 6.1 (e)) or processing is necessary for scientific research purposes in the public interest and in accordance with Article 89 (1) - (Article 9.2(j)) for data collected from NSIs, data about respondents’ neighbourhood/area, and reasons for not participating etc. However, this legal basis is subject to national implementation and may vary between countries.

If processing is based on Article 6, 1 (f) - legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject, one must describe the legitimate interest pursued by the controller;

The recipients or categories of recipients of the personal data, if any;

- Describe who will have access to contact information and indirectly identifiable data

Where applicable, the fact that the controller intends to transfer personal data to a third country (meaning outside EU/EEA) or international organisation;

- This could be relevant if any of the processors stores their data outside of EU/EEA, e.g. by using cloud service providers as sub processors.
- Or e.g. if researchers from outside EU/EEA will have access to non-anonymous raw data.

The period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;

- For how long will contact information be stored?
- For how long will non-anonymous raw data be stored?

The existence of the data subject’s rights;

- Access to, rectification or erasure of personal data
- Restriction of processing
- Objection to processing
- Data portability

- The right to withdraw consent at any time without affecting the lawfulness of processing based on consent before its withdrawal
    - If the controller or processor is no longer capable of re-identifying the data subject after the contact information/scrambling key is deleted without collecting additional information, the controller may define that data subjects are free to withdraw (and thus require their personal data to be deleted) up until this point (cf. Article 11(1))

The right of participants to lodge a complaint with a supervisory authority;

Information on any processing of the personal data for a purpose other than that for which the personal data were collected;

Article 14 - Additional information to provide when data is collected from third parties

For data not collected directly from the data subjects, the controller shall in addition provide the data subject with the categories of these personal data (Article 14(d)). Information must be provided about all types of data that will be collected from third parties. The categories of these auxiliary data, from which sources they come from and (if applicable) whether they are publicly available must be included. Concerns e.g.:

- Information from the sampling frame
- Collection of neighbourhood contextual information
- Interviewer's notes of the non-respondent's reasons not to participate, like language barriers, illness etc
- Planned future linkage to admin/other auxiliary data

Withdrawal of consent

Withdrawal of consent is given a prominent place in the GDPR. Article 7(3) prescribes that the controller must ensure that consent can be withdrawn by the data subject as easy as giving consent and at any given time. If there is no other lawful basis justifying the processing (e.g. further storage) of the data, they should be deleted or anonymised by the controller (WP29, p. 22).

The WP29 guidelines emphasize the importance of controllers assessing the purposes for which data is actually processed and the lawful grounds on which it is based prior to collecting the data. The processing could be based on more than one lawful basis. However, the application of one or more bases must be established prior to the processing and in relation to a specific purpose (WP29, p. 22).

A topic the WP29 guidelines doesn't comments on, is the relationship between Article 7(3); the right to withdraw, and Article 11(1), which calls for controllers – when the processing of personal data no longer requires the identification of a data subject – to be no longer obliged to maintain or acquire additional information to identify the data subject for the sole purpose of complying with this regulation (cf. FSD comments). The FSD exemplifies the need for further clarification with a situation from survey research:

A survey research is conducted on behalf of the controller by a company acting in the capacity of a processor. The legal basis for the data processing is data subject's consent. The rights and duties between the controller and the processor are stipulated in a contract that fulfils the criteria of Article 28. As per the contract, the processor delivers to the

controller data that is void of direct identifiers. The processor's activity ends, and no further information is retained or delivered. The data possessed by the controller no longer permits withdrawal of consent as the data subjects can no longer be identified with adequate certainty. Is consent as a legal basis invalid or does Article 11(1) apply to the situation?

Records of consent

The controller must be able to demonstrate that consent has been obtained (Article 7(1)). The WP29 guidelines on consent points out that (cf. WP29 p. 20):

- Controllers are free to develop methods that are fitting in their daily operations
- The requirement to demonstrate consent should not lead to excessive additional data processing
- The GDPR does not prescribe how this should be done in detail
- The controller must be able to prove consent *in a given case*

The WP29 guidelines give flexibility for complying with Article 7(1) without giving explicit guidance on what is an adequate level of documentation on consent. It is unclear how "methods that are fitting in their daily operations" is to be interpreted, especially in the context of scientific research. Would for instance the following practice fulfil the requirements of Article 7(1)?

A research project conducts face to face survey interviews with research participants. Information is given in advance of the interview, and time and date for the interview is settled. When the interview takes place, the interviewer first asks the participant if the information is read and understood, if the participant has any questions, and accordingly, if he/she gives their consent to participating in the research project. After this assurance, the interviewer ticks off that the participant has received information and has given his/her consent (orally) together with time and date for the consent in his/her record. The record is linked to the identity of the participant by pseudonymisation (cf. FSD comments).

Definitions of anonymisation versus pseudonymisation

Recital 26 sets out that the GDPR does not apply to data that "does not relate to an identified or identifiable natural person or to data rendered anonymous in such a way that the data subject is not or no longer identifiable". The recital reads further: "personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information, should be considered to be information on an identifiable natural person".

Then, recital 26 sets out that "to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments" (Recital 26).

Pseudonymisation is defined in Article 4(5), as the handling of personal data in such a way that no individuals can be identified from the data without a "key" that allows the data to be re-identified. It is a condition for the data to be pseudonymised that the key is kept separately and secure. Hence, this process involves removing or obscuring direct identifiers

and, in some cases, certain indirect identifiers that could combine to reveal a person's identity.

The difference between anonymisation versus pseudonymisation thus rests on whether the data can be re-identified. While anonymised data falls outside the scope of the GDPR, controllers can at least benefit from relaxed standards under the GDPR by rendering data pseudonymous (Wes 2017, Achatz 2017).

The GDPR recognises that pseudonymisation can reduce risks to the data subjects concerned and help controllers and processors meet their data-protection obligations (Article 28). Accordingly, the GDPR creates significant incentives for controllers to pseudonymise personal data. Under the GDPR, pseudonymisation can help a controller: (1) fulfill its data security obligations; (2) safeguard personal data for scientific, historical, and statistical purposes; and (3) mitigate its breach notification obligations (Achatz 2017).

Are pseudonymised data always personal data?

The Information Commissioner's Office is advising on its website that "personal data that has been pseudonymised, e.g. key-coded, can fall within the scope of the GDPR depending on how difficult it is to attribute the pseudonym to a particular individual".

Following this reservation, it is to say that pseudonymised data *can* fall within the scope of the GDPR, i.e. they *can* be personal data, but that this is not necessarily the case. Conversely, if the *can* fall outside the scope of the GDPR, then it must follow that they *can* be anonymous (Mourby et.al 2018:3).

How does this interpretation (or guidance) make sense, in light of the GDPR's definition that pseudonymous data should be considered as information on an identifiable natural person – implying all pseudonymous data - whoever they are held by?

In an early phase of the GDPR, ICO amongst others warned that recital 26, by including a definition of pseudonymisation in the context of the definition of personal data, would create unnecessary confusion. The ICO stated that instead of giving the impression that pseudonymisation is determinative of identifiability, it should rather be understood as a process to be applied to personal data as defined in Article 4(5) (Mourby et.al 2018:4).

Following up on ICO's interpretation, Mourby et.al argue that the definition of pseudonymisation under the GDPR is not intended to determine whether data are personal, but rather, that recital 26 and its requirement of "means reasonably likely to be used" remains the relevant test as to whether data are personal. This leaves open the possibility that data which have been pseudonymised still can be rendered anonymous, they argue (ibid).

The following example set out by Mourby et.al illustrates the ambiguity of the relationship between pseudonymisation and anonymization:

Pseudonomization v Anonymization

Public Authority A provides administrative personal data to a Research Centre, B, to be used for research purposes. The Research Centre wishes to share this data with Researcher C but is not sure whether they would be disclosing personal data. Research Centre B processes the personal data and removes the information which is deemed to be directly identifying. These identifiers are held separately within Research Centre B, with technical and organisational controls to prevent their reattribution to the research data. Researcher C accesses the research data in a secure lab (Research Centre B). She has completed the Centre's accreditation training so she knows she cannot bring a phone or tablet into the room where she is working on the data, and the computer she works on is not networked. In addition, she signs an agreement with Research Centre B not to attempt to identify any natural person within the data (she is interested solely in the patterns within the data, which might help their project). All her analytical outputs from her work are checked before she is allowed to take them out of the centre. Researcher C has no relationship with Research Centre B, or with Public Authority A, which would enable her to access any potentially identifying information. She has no means by which she is reasonably likely (or indeed likely) to obtain the information which would identify the data. The information exists, and so identification is not theoretically impossible, and the processing is therefore not technically irreversible. However, it is extremely unlikely that the researcher could or would have access to any information which would enable her to identify natural persons within the data. Is Researcher C accessing personal data?

Mourby et.al concludes that as it is Recital 26 of the GDPR, and not Article 4(5), which determines whether the data are personal data, this leaves open the possibility that data which have undergone pseudonymisation could be anonymous for a third party such as Researcher C (Mourby et.al 2018:4).

Indirectly identifiable data

Indirect identifiers are data that do not identify an individual in isolation but may reveal individual identities if combined with additional data points (Maldoff 2016). Within survey research, a respondent can, for instance, be indirectly identified through a combination of demographic variables such as detailed geographic information (as in state, county, or province), organisations to which the respondent belongs, educational institutions (from which the respondent graduated and year of graduation), detailed occupational titles, place where respondent grew up, exact dates of events (birth, death, marriage, divorce), detailed income, etc. As record linkage between survey data and administrative data adds more detailed background information, the risk of indirect identification in a dataset increases. Longitudinal data, with information on longitudinal links and patterns, present an additional challenge (Cox et.al 2011). A well-known example to illustrate how easily an individual can be identified, is the calculation of, based on year 2000 census data, that 63% of the US population are uniquely identifiable by ZIP code, birthdate, and gender (Golle 2006, Ohm 2010).

Article 29 Working Party (WP29) guidelines on Anonymisation Techniques from 2014 identifies seven techniques that can be used to anonymise records of information:

1. Noise Addition: The personal identifiers are expressed imprecisely (e., weight is expressed inaccurately +/- 10 lb).

2. Substitution/Permutation: The personal identifiers are shuffled within a table or replaced with random values (e., a zip code of 80629 is replaced with “Magenta”).
3. Differential Privacy: The personal identifiers of one data set are compared against an anonymized data set held by a third party with instructions of the noise function and acceptable amount of data leakage.
4. Aggregation/K-Anonymity: The personal identifiers are generalized into a range or group (e., a salary of $42,000 is generalized to $35,000 - $45,000).
5. L-Diversity: The personal identifiers are first generalized, then each attribute within an equivalence class is made to occur at least “l” times. (e., properties are assigned to personal identifiers, and each property is made to occur with a dataset, or partition, a minimum number of times).
6. Pseudonymization – Hash Functions: The personal identifiers of any size are replaced with artificial codes of a fixed size (e., Paris is replaced with “01”, London is replaced with “02”, and Rome is replaced with “03”).
7. Pseudonymization – Tokenization: The personal identifiers are replaced with a non-sensitive identifier that traces back to the original data, but are not mathematically derived from the original data (i.e., a credit card number is exchanged in a token vault with a randomly generated token “958392038”).

4 Conclusions and Recommendations

The potential advantages of administrative data linkage are well documented (Connelly et al., 2016). One of the most potent of these is the ability of researchers to bring their analysis and conclusions closer to the processes of government and society and potentially include the societal relevance of social research generally. Yet despite this promise administrative data linkage is not pervasive within social science research and the proportion of peer-reviewed studies which draw on administrative data as part of their empirical base is still very low.

In this report we have examined the various infrastructure and initiatives aimed at facilitating social science research through linked administrative data. Generally, these have serious limitations. When assessing administrative data along the FAIR principles it performs exceptionally poorly. It is not readily findable with metadata of administrative records being rarely searchable by the scientific community. It is not very accessible as there are strong and often opaque barriers to accessing such data that go beyond data protection regulations. Even in instances where significant investment is made in supporting and facilitating access such as through the Administrative Data Research Network in the UK, there remain serious barriers to access as such investments are not made in data controllers.

Administrative data is also rarely interoperable. The use of persistent identifiers in survey research is increasingly common but many administrative data systems have poor management of such persistent identifiers leading to mismatches, duplications or lost records (Abowd et al., 2018). To support the quality of these persistent identifiers and their use in administrative data linkage, it is good practice to use the administrative data records as a sampling frame for the survey in the first instance. This enables such quality issues to be incorporated and accounted for in the survey design. Achieving this will however require social science infrastructures to actively lobby for better sampling frames and sampling access such as the activities in work package 2 of SERISS.

Finally, administrative data needs to be reusable. The use of tools such as Jupyter notebooks needs to be central within administrative data infrastructures and access procedures need to be constructed in such a way that inherently allows for replication studies. This is not common in administrative data linkage studies but should become best practice. There is nothing within data protection regulations that would necessarily prevent this as the purpose of processing would remain the same.

Nevertheless, there are many reasons to be optimistic about the prospects of linked administrative data within social science research. A shared and clear legal framework provided for by the General Data Protection Regulation which explicitly allows for processing for research purposes is a significant improvement for social science research, particularly in the acquisition of sampling frames which are the basis on which future administrative data linkage can be conducted. From a technical perspective, there are also a wide range of possibilities that are opening up with regards to distributed computation and secure cloud services which could further expand researchers' opportunities for linking to administrative data. In the context of the European Open Science Cloud, improved e-infrastructures will be vital to this end and will be a goal shared across a number of scientific disciplines including the health and biomedical sciences.

The final conclusion of this report is however a note of caution which is drawn from the experiences of all the infrastructures and initiatives that are included here. For administrative data linkage to be feasible for research purposes, it must include the 'buy-in' of data controllers. In the ADRN, this was not evident and led to many stalled projects. In projects such as IDAN, ODISSEI or CASD, this was more evident. Social Science Infrastructures aimed at administrative data linkage should think carefully about what they can offer such data controllers. This could include training in advanced analytic techniques, improved computational hardware or innovative data sources. Regardless, administrative data infrastructures must have administrative data controllers as central stakeholders and an operational and business model that clearly serves them in a unique and constructive way.

References

Abowd, J. M., McKinney, K. L., & Zhao, N. L. (2018). Earnings inequality and mobility trends in the United States: Nationally representative estimates from longitudinally linked employer-employee data. Journal of Labor Economics, 36(S1), S183-S300. https://doi.org/10.1086/694104/

Administrative Data Research Network – ADRN (2017), "Frequently Asked Questions – About the Data", accessed via https://adrn.ac.uk/for-the-public/faq/?About-the-data.

Andersson, Gunnar, Michaela Kreyenfeld and Tatjana Mika (2014), "Welfare state context, female labour-market attachment and childbearing in Germany and Denmark" *Journal of Population research* 31: 287-316.

Bach, Mirjana, Andjelko Lojpur, Sanja Pekovic and Tatjana Stanovcic (2015), "The influence of different information sources on innovation performance: evidence from France, the Netherlands and Croatia" *South East European Journal of Economics and Business* 10(2): 89-101.

Bingley, Paul and Alessandro Martinello (2014), "Measurement error in the Survey of Health, Ageing and Retirement in Europe: a validation study with administrative data for educational

level, income and employment". SHARE Working Paper Series, n. 16-2014, accessed via http://www.share-project.org/uploads/tx_sharepublications/WP_Series_16_2014_Bingley_Martinello_01.pdf.

Bond, Steve, Maurice Brandt and Peter-Paul de Wolf (2015), "Guidelines for output checking". Data without Borders, accessed via http://www.dwbproject.org/access/guides.html.

Börsch-Supan, Axel, Benedikt Alt and Tabea Bucher-Koenen (2015), "Early retirement for the underpriviledged? Using the record-linked SHARE-RV data to evaluate the most recent German pension reform". In *Ageing in Europe - Supporting Policies for an Inclusive Society*, ed. by Axel Börsch-Supan, Thorsten Kneip, Howard Litwin, Michal Myck and Gugliemo Weber. Berlin: De Gruyter.

Brass, William (1975), *Methods for Estimating Fertility and Mortality from Limited and Defective Data*. Chapel Hill: Carolina Population Centre.

Buber-Ennser, Isabella (2014), "Attrition in the Austrian Generations and Gender Survey: is there a bias by fertility relevant aspects?" *Demographic Research* 31: 459-496.

Butt, Sarah and Kaisa Lahtinen (2016), "ADDResponse: Auxiliary Data Driven nonResponse Bias Analysis. Technical report on appending geocoded auxiliary data to Round 6 of European Social Survey (UK)", accessed via http://doc.ukdataservice.ac.uk/doc/8066/mrdoc/pdf/8066_addresponse_technical_report_final.pdf

Butt, Sarah and Kaisa Lahtinen (2016a), "Getting to grips with different types of auxiliary data: was it worth it?", presented at the Workshop 'Tackling survey nonresponse: the role of geocoded auxiliary data", accessed via https://blogs.city.ac.uk/addresponse/files/2016/03/ButtLahtinen_Final-2bpf86s.pdf.

Brewer, Michael and Jonathan Cribb (2016), "Lone Parents, Time-limited in-work credits and the dynamics of work and welfare" The Institute of Labor Economics Discussion Paper Series, n. 10414, accessed via http://ftp.iza.org/dp10414.pdf.

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

Card, David, Raj Chetty, Martin Feldstein and Emmanuel Saez (2010), "Expanding access to administrative data for research in the United States". Paper written for the National Science Foundation 10-069 call on the "Future Research in the Social, Behavioural and Economic Sciences", accessed via https://eml.berkeley.edu/~saez/card-chetty-feldstein-saezNSF10dataaccess.pdf.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach and Danny Yagan (2011), "How does your kindergarten classroom affect your earnings? Evidence from Project STAR" *Quarterly Journal of Economics* 127(4): 1593-1660.

Eichenlaub, Suzanne, Stewart Tolnay and J. Trent Alexander (2010), "Moving out but not up: economic outcomes in the great migration" *American Sociological Review* 75(1): 101-125.

Emery, Tom (2016), “Administrative data linking: enriching administrative data with surveys”. Paper presented at the European Conference on Quality in Official Statistics, 31 May-3 June 2016, Madrid.

Esteve, Albert, Joan Garcia-Roman and Ron Lesthaeghe (2012), “The family context of cohabitacion and single motherhood in Latin America” *Population and Development Review* 38(4): 707-727.

Esteve, Albert (2011), “Trans-border access to census microdata: the IPUMS-IECM partnership, where a single license agreement opens access to microdata for more than 60 countries to researchers world-wide free of cost”. Paper presented at the Joint UNECE/ Eurostat work session on statistical data confidentiality, Tarragona.

Groenewold, George, Jeroen van Ginneken and Cristina Masseria (2008). “Towards comparable statistics on mortality by socioeconomic status in EU member states. Methodology note” European Commission. Accessed via http://ec.europa.eu/social/BlobServlet?docId=3958&langId=en.

Gurke, Christopher, David Schiller, Kamel Gadouche, Steve Bond and Atle Alvheim (2012), “Report on the state of the art of current SC in Europe”, Data without Borders (deliverable D4.1), accessed via http://www.dwbproject.org/export/sites/default/about/public_deliveraples/d4_1_current_sc_in_europe_report_full.pdf.

Hill, Kenneth (1987), “Estimating census and death registration completeness” *Asian and Pacific Population Forum* 1(3): 8-13.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Nordholt, Keith Spicer and Peter-Paul de Wolf (2012), *Statistical Disclosure Control*. New York: Wiley.

Itoh, Shinsuke and Masahiro Takano (2011). “A method to quantitatively assess confidentiality and potential usage of official microdata in Japan”. Paper presented at the Proceedings of the 58th World Statistical Congress, 2011, Session CPS020, Dublin.

Karadja, M., Mollerstrom, J., & Seim, D. (2017). Richer (And Holier) Than Thou? The Effect of Relative Income Improvements on Demand for Redistribution. Review of Economics & Statistics, 99(2), 201–212. https://doi.org/10.1162/REST_a_00623

Künn, Steffen (2015). “The challenges of linking survey and administrative data.” *IZA World of Labor 214*. doi: 10.15185/izawol.214, accessed via http://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data.pdf.

Lahtinen, Kaisa and Sarah Butt (2016), “Understanding nonresponse behavior on the European Social Survey: the role of survey paradata vs. external auxiliary data”. Paper presented at the 7th ESRC Research Methods Festival, Bath, accessed via https://blogs.city.ac.uk/addresponse/files/2016/08/ADDResponse-RMF-July016_v2pptx-2insj7r.pdf.

Lam, David and Leticia Marteleto (2008), “Stages of the demographic transition from a child’s perspective” *Population and Development Review* 34(2): 225-252.

Lifang, Gu, Rohan Baxter, Deanne Vickers and Chris Rainsford (2003), *Record Linkage: Current Practice and Future Directions*. CMIS Technical Report 03/83. Accessed via http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.8119.

Longhi, Simonetta (2013), "Impact of cultural diversity on wages, evidence from panel data" *Regional Science and Urban Economics* 43: 797-807.

Madsen, Peter (2003), "The Ethics of Confidentiality: The Tension Between Confidentiality and the Integrity of Data Analysis in Social Science Research". NSF Workshop on Confidentiality Research. Arlington, Virginia.

McCaa, Robert and Albert Esteve (2012), "Disclosure controls for one-stop, trans-border access to census microdata for 98 countries via a single license and dissemination point: the IPUMS-IECM partnership". Paper presented at the ESSnet workshop on Statistical Disclosure Control of Census data, Luxemburg.

McGrath-Lone, Louise, Lorraine Dearden, Katie Harron, Bilal Nasim and Ruth Gilbert (2016), "Factors associated with re-entry to out-of-home care among children in England" *Child Abuse and Neglect* 63: 73-83.

OECD (2005), "Statistical disclosure control", Glossary of Statistical Terms, accessed via https://stats.oecd.org/glossary/detail.asp?ID=6996.

Playford, Christopher J., Vernon Gayle, Roxanne Connelly, and Alasdair JG Gray. "Administrative social science data: The challenge of reproducible research." Big Data & Society 3, no. 2 (2016)

Poulain, Michel and Anne Herm (2013), "Central population registers as a source of demographic statistics in Europe" *Population* 68(2): 183-212, accessed via http://www.cairn-int.info/article-E_POPU_1302_0215--central-population-registers-as-a-source.htm.

Rasner, Anika, Joachim Frick and Markus Grabka (2013), "Statistical matching of administrative and survey data: an application to wealth inequality analysis" *Sociological Methods and Research* 42(2): 192-224.

Regulation of the European Parliament and European Council 2016/679 of 27 April 2016 on the *Protection of natural persons with regard to the processing of natural data and on the free movement of such data* (OJ 119, 4/5/2016).

Ruggles, Steven, Robert McCaa, Matthew Sobek and Lara Cleveland (2015), "The IPUMS collaboration: integrating and disseminating the world's population microdata" *Journal of Demographic Economics* 81(2): 203-216.

Sayers, Adrian, Yoav Ben-Shlomo, Ashley Blom and Fiona Steele (2015). "Probabilistic record linkage". *International Journal of Epidemiology*. doi: 10.1093/ije/.

Schaake, Karina, Jack Burgers and Clara Mulder (2013), "Ethnicity, education and income, and residential mobility between neighbourhoods". *Journal of Ethnic and Migration Studies* 40(4): 512-527.

SHARE (2017), "User Guide Release 6.0.0", accessed via http://www.share-project.org/fileadmin/pdf_documentation/User_Guide_6-0-0.pdf.

Silberman, Roxane, Jara Kampmann, Maurice Brandt, Eric Debonnel, Katharina Kinder-Kurlanda, Mathias Zenke, Philippe Donnay, Kamel Gadouche (2015), “Proof of concept for a European network of secure remote access systems”, presented at European Data Access Forum, Luxemburg.

Stoop, Ineke, Jaak Billiet, Achim Koch and Rory Fitzgerald (2010), *Improving Survey Response. Lessons Learned from the European Social Survey.* Chichester: Wiley.

Sutherland, Alex, Sonia Ilie and Anna Vignoles (2015), “Factors associated with achievement: key stage 4. Research Report”. Department for Education. Accessed via https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/473673/RR407_-_Factors_associated_with_achievement_-_key_stage_4.pdf.

United Nations Economic Commission for Europe (2007), *Managing statistical confidentiality and microdata access. Principles and guidelines of good practice.* New York and Geneva: United Nations Publications.

United Nations Economic Commission of Europe (2007), “Register-based statistics in the Nordic Countries. Review of best practices with focus on population and social statistics”. Accessed via http://www.unece.org/index.php?id=17470.

Zueras, Pilar and Pau Miret Gamundi (2013), “Elderly Who Live Alone: An Overview based on the 1991 and 2001 Censuses” *Revista Española de Investigaciones Sociológicas* 144: 139-152.

Annex 1: Annotated bibliography

Abowd, J. M., & Stinson, M. H. (2013). Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data. *Review of Economics and Statistics, 95*(5), 1451-1467. doi: 10.1162/Rest_a_00352

We propose a new methodology that does not assume a priori specification of the statistical properties of the measurement errors and treats all sources as noisy measures of some underlying true value. The unobservable true value can be represented as a weighted average of all available measures, using weights that must be specified a priori unless there has been a truth audit. The Census Bureau's Survey of Income and Program Participation (SIPP) survey jobs are linked to Social Security Administration earnings data, creating two potential annual earnings observations. The reliability statistics for both sources are quite similar except for cases where the SIPP used imputations for some missing monthly earnings reports.

Abraham, J. M., Karaca-Mandic, P., & Boudreaux, M. (2013). Sizing Up the Individual Market for Health Insurance: A Comparison of Survey and Administrative Data Sources. *Medical Care Research and Review, 70*(4), 418-433.

Provisions within the Affordable Care Act, including the introduction of subsidized, exchange-based coverage for lower income Americans lacking access to employer coverage, are expected to greatly expand the size and importance of the individual market. Using multiple federal surveys and administrative data from the National Association of Insurance Commissioners, we generate national-, regional-, and state-level estimates of the individual market. In 2009, the number of nonelderly persons with individual coverage ranged from 9.55 million in the Medical Expenditure Panel Survey to 25.3 million in the American Community Survey. Notable differences also exist between survey estimates and National Association of Insurance Commissioners administrative counts, an outcome likely driven by variation in the type and measurement of individual coverage considered by surveys relative to administrative data. Future research evaluating the impact of the Affordable Care Act coverage provisions must be mindful of differences across surveys and administrative sources as it relates to the measurement of individual market coverage.

Aldhouse, F. (2013). Data protection in Europe - Some thoughts on reading the academic manifesto. *Computer Law & Security Review, 29*(3), 289-292. doi: 10.1016/j.clsr.2013.03.013

The commentary by academics on the proposed European General Data Protection Regulation in [2013] 29 CLSR 180 has provoked thoughts in response. The responder strongly agrees with the doubts expressed about the definition of personal data, anonymisation and the identifiability of individuals. On the other hand, he disagrees with the views on consent and legitimacy and proposes support for a risk-based approach to data protection. He suggests that data protection does not need to be defended from the attack that it stifles business, but is justifiable for its assertion of fundamental rights. In conclusion, he shares the criticism of the European Commission's delegated and implementing powers and is concerned that the Regulation will be rushed to a conclusion for reasons of political ambition. (C) 2013 Francis Aldhouse. Published by Elsevier Ltd. All rights reserved.

Aldhouse, F. (2014). Anonymisation of personal data - A missed opportunity for the European Commission. *Computer Law & Security Review, 30*(4), 403-418.

As early as the 1970's, privacy studies recognised that 'anonymization' needed to be approached with caution. This caution has since been vindicated by the increasing sophistication of techniques for re-identification. Yet the courts in the UK have so far only hesitatingly grappled with the issues involved, while European courts have produced no guidance. Reviewing the limited case law, the author finds the concepts of both 'personal data' (which must be protected) and 'anonymization' (which removes this requirement) misleadingly simplistic. A more practical approach would recognise that identifiability sits on a continuum so that regulation needs to be risk-based and proportional. He proposes some consequential changes to the proposed EU Regulation, albeit with modest hopes for success. This paper is a

shortened and slightly revised version of a dissertation submitted in April 2013 to Staffordshire University for the award of the degree of LLM.

Bagger, J., & Seltzer, A. (2014). Administrative and Survey Data in Personnel Economics. *Australian Economic Review, 47*(1), 137-146. doi: 10.1111/1467-8462.12051

The number of empirical studies in personnel economics using administrative and survey data has grown rapidly in recent years. We survey the use of administrative data to examine employment contracts. Specifically, we consider three types of data that have been widely used in empirical studies: historical firm-level records, contemporary firm-level records and national matched employer-employee records. Studies using this sort of administrative data have shed considerable light on the nature of employment relationships.

Beebe, T. J., Ziegenfuss, J. Y., Jenkins, S. M., Haas, L. R., & Davern, M. E. (2011). Who Doesn't Authorize the Linking of Survey and Administrative Health Data? A General Population-based Investigation. *Annals of Epidemiology, 21*(9), 706-709.

PURPOSE: To determine the extent of authorization bias in a study linking survey and medical record dam in a general population-based investigation. METHODS: Authorization status (authorized data linkage was ascertained through a sequential mixed mode mail and telephone survey conducted in Olmsted County, MN. Respondents (regardless of authorization status) were linked to the Rochester Epidemiology Project (REP), the medical record system; for health care providers in Olmsted County. The REP provided data on gender, age, race, health status (comorbid conditions) and health care utilization (ER admission. hospital admission, clinical office visits and procedures). Authorizers (n = 1357) are compared to non-authorizers (n=217) with respect to these demographic and clinical characteristics. RESULTS: 86.2% of respondents authorized data linkage. Non-authorizers were younger, healthier (lower Charlson score), and less likely to have 3 or more recent clinical office visits. In multivariate analysis, Charlson score was no longer a significant predictor of authorization while an ER visit did predict: authorization. CONCLUSIONS: Younger subjects are less likely to authorize data linkages. As such researchers should be aware of this source of potent potential bias when analysing population-based linked survey and administrative data. The presence of hits with respect to health care users is more complicated. It is dependent on how the concept is operationalized with heavy clinical users more likely to authorize and those with ER visits less so. Ann I epidemiology 2011;21:706-709.

Blais, C., Rochette, L., Hamel, D., & Poirier, P. (2014). Prevalence, incidence, awareness and control of hypertension in the province of Quebec: Perspective from administrative and survey data. *Canadian Journal of Public Health-Revue Canadienne De Sante Publique, 105*(1), E79-E85.

OBJECTIVES: Hypertension is a major risk factor for cardiovascular diseases. Nearly one adult in four was diagnosed with hypertension in 2007-2008 in Canada. One of the objectives of this study was to determine whether the prevalence of hypertension in Quebec as assessed using administrative data is comparable to that specifically measured in surveys, especially for the elderly population. METHODS: Trends in prevalence, incidence and mortality were examined using the Quebec Integrated Chronic Disease Surveillance System built from grouping numerous administrative databases from 1996-1997 to 2009-2010. Blood pressure measurements, hypertension prevalence, awareness and control were obtained in 1,706 Quebecers in the combined cycles of the Canadian Health Measures Survey. RESULTS: Using administrative databases, 23.6% [95% confidence interval, 23.5-23.6] of the Quebec population (n=1,433,400) aged >= 20 years was diagnosed with hypertension in 2009-2010, an increase of 32.1% compared to 2000-2001. The incidence decreased by 27.3%. Among people aged >= 65 years, the prevalence rose to 69.0% [95% CI: 68.8-69.2] in women and 61.7% [95% CI: 61.5-61.9] in men. For people aged 20-79 years, the prevalence of hypertension was lower with the administrative data compared to the survey (20.2% and 23.1%, respectively). The level of awareness, treatment and control were 84.3%, 83.1% and 67.9%, respectively. CONCLUSION: The prevalence of hypertension derived from administrative data is comparable to that obtained with a health measured survey. Elderly women (>= 65 years) are a very high-risk subgroup. The levels of awareness, treatment and control of hypertension in Quebec are very high.

Brass W. (1975). Methods for Estimating Fertility and Mortality from Limited and Defective Data. Chapel Hill: Carolina Population Centre.

In many developing countries, vital registration systems are incomplete, and demographic data collected by other methods, such as censuses and surveys, are defective. After a brief account of different strategies for coping with this situation, there is a description of Brass's general principles for obtaining estimates from unreliable data. These principles are: (i) serendipity, or making the most of one's chances; (ii) rehabilitation, or interfering as little as possible; (iii) consistency, or making use of underlying demographic relationships; (iv) robustness, that good estimates can be made even when the underlying assumptions are not fully met; and (v) no rule, or nothing works all the time. Four useful general devices, cumulation, the use of reference standards, linearizing transformations, and scaling transformations, are then illustrated. All the techniques described in this volume have already been published elsewhere, although often in obscure and hard to obtain United Nations reports. The great value of this collection is that all the important methods developed by Professor Brass for dealing with defective or deficient data have been brought together in one place, with examples of their application, and an explanation of the principles of his general approach to data adjustment.

Brion, P., & Gros, E. (2015). Statistical Estimators Using Jointly Administrative and Survey Data to Produce French Structural Business Statistics. *Journal of Official Statistics, 31*(4), 589-609.

Using as much administrative data as possible is a general trend among most national statistical institutes. Different kinds of administrative sources, from tax authorities or other administrative bodies, are very helpful material in the production of business statistics. However, these sources often have to be completed by information collected through statistical surveys. This article describes the way Insee has implemented such a strategy in order to produce French structural business statistics. The originality of the French procedure is that administrative and survey variables are used jointly for the same enterprises, unlike the majority of multisource systems, in which the two kinds of sources generally complement each other for different categories of units. The idea is to use, as much as possible, the richness of the administrative sources combined with the timeliness of a survey, even if the latter is conducted only on a sample of enterprises. One main issue is the classification of enterprises within the NACE nomenclature, which is a cornerstone variable in producing the breakdown of the results by industry. At a given date, two values of the corresponding code may coexist: the value of the register, not necessarily up to date, and the value resulting from the data collected via the survey, but only from a sample of enterprises. Using all this information together requires the implementation of specific statistical estimators combining some properties of the difference estimators with calibration techniques. This article presents these estimators, as well as their statistical properties, and compares them with those of other methods.

Bulloch, A. G., Currie, S., Guyn, L., Williams, J. V., Lavorato, D. H., & Patten, S. B. (2011). Estimates of the treated prevalence of bipolar disorders by mental health services in the general population: comparison of results from administrative and health survey data. *Chronic Diseases and Injuries in Canada, 31*(3), 129-134.

Introduction: Informed provision of population mental health services requires accurate estimates of disease burden. Methods: We estimated the treated prevalence of bipolar disorders by mental health services in the Calgary Zone, a catchment area in Alberta with a population of over one million. Administrative data in a central repository provides information of mental health care contacts for about 95% of publically funded mental health services. We compared this treated prevalence against self-reported data in the 2002 Canadian Community Health Survey: Mental Health and Well-Being (CCHS 1.2). Results: Of the 63 016 individuals aged 18 years plus treated in the Calgary Zone in 2002-2008, 3659 (5.81%) and 1065 (1.70%) were diagnosed with bipolar I and bipolar II disorder, respectively. The estimated treated population prevalence of these disorders was 0.41% and 0.12%, respectively. We estimated that 0.44% to 1.17% of the Canadian population was being treated by psychiatrists for bipolar I disorder from CCHS 1.2.Discussion: For bipolar I disorder the estimate based on local administrative data is close to the lower end of the health survey range. The degree of agreement in our estimates reinforces the utility of administrative data repositories in the surveillance of chronic mental disorders.

Bustard, J. (2015). The Impact of EU Privacy Legislation on Biometric System Deployment [Protecting citizens but constraining applications]. *Ieee Signal Processing Magazine, 32*(5), 101-108. doi: 10.1109/Msp.2015.2426682

Biometric systems provide a valuable service in helping to identify individuals from their stored personal details. Unfortunately, with the rapidly increasing use of such systems, there is a growing concern about the possible misuse of that information. To counteract the threat, the European Union (EU) has introduced comprehensive legislation that seeks to regulate data collection and help strengthen an individual's right to privacy. This article looks at the implications of the legislation for biometric system deployment. After an initial consideration of current privacy concerns, the definition of personal data and its protection is examined in legislative terms. Also covered are the issues surrounding the storage of biometric data, including its accuracy, its security, and justification for what is collected. Finally, the privacy issues are illustrated through three biometric use cases: border security, online bank access control, and customer profiling in stores.

Carinci, F. (2015). Essential levels of health information in Europe: An action plan for a coherent and sustainable infrastructure. Health Policy, 119(4), 530-538.

The European Union needs a common health information infrastructure to support policy and governance on a routine basis. A stream of initiatives conducted in Europe during the last decade resulted into several success stories, but did not specify a unified framework that could be broadly implemented on a continental level. The recent debate raised a potential controversy on the different roles and responsibilities of policy makers vs the public health community in the construction of such a pan-European health information system. While institutional bodies shall clarify the statutory conditions under which such an endeavour is to be carried out, researchers should define a common framework for optimal cross-border information exchange. This paper conceptualizes a general solution emerging from past experiences, introducing a governance structure and overarching framework that can be realized through four main action lines, underpinned by the key principle of "Essential Levels of Health Information" for Europe. The proposed information model is amenable to be applied in a consistent manner at both national and EU level. If realized, the four action lines outlined here will allow developing a EU health information infrastructure that would effectively integrate best practices emerging from EU public health initiatives, including projects and joint actions carried out during the last ten years. The proposed approach adds new content to the ongoing debate on the future activity of the European Commission in the area of health information.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding Access to Administrative Data for Research in the United States. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas. http://ssrn.com/abstract=1888586 or http://dx.doi.org/10.2139/ssrn.1888586

We argue that the development and expansion of direct, secure access to administrative micro-data should be a top priority for the NSF. Administrative data offer much larger sample sizes and have far fewer problems with attrition, non-response, and measurement error than traditional survey data sources. Administrative data are therefore critical for cutting-edge empirical research, and particularly for credible public policy evaluation. Although a number of agencies have successful programs to provide access to administrative data - most notably the Centers for Medicare and Medicaid Services - the United States generally lags far behind other countries in making data available to researchers. We discuss the value of administrative data using examples from recent research in the United States and abroad. We then outline a plan to develop incentives for agencies to broaden data access for scientific research based on competition, transparency, and rewards for producing socially valuable scientific output.

Celik, S., Juhn, C., McCue, K., & Thompson, J. (2012). Recent Trends in Earnings Volatility: Evidence from Survey and Administrative Data. *B E Journal of Economic Analysis & Policy, 12*(2). doi: Artn 110.1515/1935-1682.3043

Recent papers find that earnings volatility is again on the rise (Dynan et al. 2008, and Shin and Solon 2011). Using household survey data-the matched Current Population Surveys and Survey of Income and Program Participation-and the newly available Longitudinal Employment

and Household Dynamics administrative dataset, we find that earnings volatility was remarkably stable in the 1990s and through the mid 2000s. This evidence is in contrast to that from the Panel Study of Income Dynamics (PSID) which registers a sharp increase in the early 2000s. We investigate whether adjusting measures based on our sources to more closely match the characteristics of the PSID can reconcile this divergence in trends, but do not find a clear explanation for the divergence. We also find little evidence of a rise over this period in the components of volatility: volatility among job leavers, volatility among job stayers, and the fraction of workers who are job leavers.

Cho, S. H., & Yun, S. C. (2009). Bed-to-nurse ratios, provision of basic nursing care, and in-hospital and 30-day mortality among acute stroke patients admitted to an intensive care unit: Cross-sectional analysis of survey and administrative data. *International Journal of Nursing Studies, 46*(8), 1092-1101. doi: 10.1016/j.ijnurstu.2009.02.001

Background: The literature reports inconsistent evidence of the effects of nurse staffing on mortality despite continuing examination of this association. Objective: To examine differences in provision of basic nursing care and in-hospital and 30-day mortality by nurse staffing of ICUs and general wards among acute stroke patients admitted to ICUs during hospitalization. Design: A cross-sectional design that included Survey and administrative data. Settings and participants: The Study included 6957 patients with haemorrhagic and ischemic stroke who were admitted to ICUs of 185 Korean hospitals. Methods: Nurse staffing of ICUs and general wards was graded based oil the bed-to-nurse ratios of each hospital. Provision of basic care was measured by whether five activities, such as bathing and feeding assistance, were fully provided by ICU nursing staff without delegation to patient families. Hospitals were categorized into low, middle, and high mortality groups for in-hospital and 30-day mortality based on z-scores that indicated standardized difference between observed and expected mortality after controlling for patient characteristics. Results: In 83.8% of hospitals, basic care was provided fully by ICU nursing staff. The overall in-hospital and 30-day mortality rates were 21.9 and 25.4%, respectively. Hospitals with higher ICU staffing were more likely to fully provide basic care. Better ICU and general staffing tended to be associated with lower in-hospital and 30-day mortality. Compared with in-hospital mortality, 30-day mortality had a more distinct increase as nurse staffing became worse. Conclusion: The findings provide evidence that nurse staffing may impact provision of basic care and patient mortality and suggest the need for policies for providing adequate nurse staffing.

Cooper, B., & Fearn, R. (1998). Dementia care needs in an area population: Case register data and morbidity survey estimates. *International Journal of Geriatric Psychiatry, 13*(8), 550-555. doi: Doi 10.1002/(Sici)1099-1166(199808)13:8<550::Aid-Gps822>3.0.Co;2-V

Objective. To compare case register data on the frequency and distribution of known dementia cases in a metropolitan area population with expected total numbers computed from a national disability survey. Method. Known cases were enumerated by a cross-sectional census of the Camberwell Dementia Register. Expected total numbers were calculated using the Cognitive Disability (CD) Planning Model, based on the OPCS national survey of disability, 1985-86. Results. Cases ascertained by the Dementia Register census comprised one-fifth of expected total prevalence. The proportion of such cases was higher for persons in long-stay care (1 in 3) than for those in private households (1 in 7). According to the CD Planning Model, cases known to specialist agencies were on average no more severely disabled and dependent than those who were unknown. In terms of absolute numbers, the district nursing and home help services appeared to be the most important untapped sources of case detection, but other research indicates that general practice contacts (not included in the planning model) may be at least equally important. Conclusions. At any given time, a high proportion of dementia cases, whether in long-stay care or in the community, will be outside the purview of specialist services. Primary care agencies are a major potential source, and a systematic health screening of persons aged over 75 years could be used to realize this potential.

Dahl, M., De Leire, T., & Schwabish, J. A. (2011). Estimates of Year-to-Year Volatility in Earnings and in Household Incomes from Administrative, Survey, and Matched Data. *Journal of Human Resources, 46*(4), 750-774.

We document trends in the volatility in earnings and household incomes between 1985 and 2005 in three different data sources: administrative earnings records, the Survey of Income and Program Participation (SIPP) matched to administrative earnings records, and SIPP survey data. In all data sources, we find a substantial amount of year-to-year volatility in workers' earnings and household incomes. In the data sources that contain administrative earnings, we find that volatility has been roughly constant, and has even declined slightly, since the mid-1980s. These findings differ from what is found using survey data and what has been reported in previous studies.

Davern, M., Call, K. T., Ziegenfuss, J., Davidson, G., Beebe, T. J., & Blewett, L. (2008). Validating health insurance coverage survey estimates - A comparison of self-reported coverage and administrative data records. *Public Opinion Quarterly, 72*(2), 241-259.

We administered a health insurance coverage survey module to a sample of 4,575 adult Blue Cross and Blue Shield of Minnesota (BCBS) members to examine if people who have health insurance coverage self-report that they are uninsured. We were also interested in whether respondents correctly classify themselves as having commercial, Medicare, MinnesotaCare, and/or Medicaid coverage (the four sample strata). The BCBS of Minnesota sample is drawn from both public and commercial health insurance coverage strata that are important to policy research involving survey data. Our findings support the validity of our health insurance module for determining whether someone who has health insurance is correctly coded as having health insurance coverage, as only 0.4 percent of the BCBS members answered the survey as though they were uninsured. However, we find problems for researchers interested in using survey responses to specific types of public coverage. For example, 21 percent of the Medicaid self-reported coverage came from known enrolees and only 67 percent of the MinnesotaCare self-reported count came from known enrolees. We conclude with a discussion of the study's implications for understanding the Medicaid "undercount" and the validity of self-reported health insurance coverage.

de Hert, P., & Papakonstantinou, V. (2016). The new General Data Protection Regulation: Still a sound system for the protection of individuals? *Computer Law & Security Review, 32*(2), 179-194.

The five-year wait is finally over; a few days before expiration of 2015 the "trilogue" that had started a few months earlier between the Commission, the Council and the Parliament suddenly bore fruit and the EU data protection reform package has finally been concluded. As planned since the beginning of this effort a Regulation, the General Data Protection Regulation is going to replace the 1995 Directive and a Directive, the Police and Criminal Justice Data Protection Directive, the 2008 Data Protection Framework Decision. In this way a long process that started as early as in 2009, peaked in early 2012, and required another three years to pass through the Parliament's and the Council's scrutiny is finished. Whether this reform package and its end-result is cause to celebrate or to lament depends on the perspective, the interests and the expectations of the beholder. Four years ago we published an article in this journal under the title "The proposed data protection Regulation replacing Directive 95/46/EC: A sound system for the protection of individuals". This paper essentially constitutes a continuation of that article: now that the General Data Protection Regulation's final provisions are at hand it is possible to present differences with the first draft prepared by the Commission, to discuss the issues raised through its law-making passage over the past few years, and to attempt to assess the effectiveness of its final provisions in relation to their declared purposes.

Drapeau, A., Boyer, R., & Diallo, F. B. (2011). Discrepancies between survey and administrative data on the use of mental health services in the general population: findings from a study conducted in Quebec. *Bmc Public Health, 11*.

Background: Population surveys and health services registers are the main source of data for the management of public health. Yet, the validity of survey data on the use of mental health services has been questioned repeatedly due to the sensitive nature of mental illness and to the risk of recall bias. The main objectives of this study were to compare data on the use of mental health services from a large scale population survey and a national health services register and to identify the factors associated with the discrepancies observed between these two sources of data. Methods: This study was based on the individual linkage of data from the

cycle 1.2 of the Canadian Community Health Survey (CCHS-1.2) and from the health services register of the Regie de l'assurance maladie du Quebec (RAMQ). The RAMQ is the governmental agency managing the Quebec national health insurance program. The analyses mostly focused on the 637 Quebecer respondents who were recorded as users of mental health services in the RAMQ and who were self-reported users or non users of these services in the CCHS-1.2. Results: Roughly 75%, of those recorded as users of mental health services users in the RAMQ's register did not report using mental health services in the CCHS-1.2. The odds of disagreement between survey and administrative data were higher in seniors, individuals with a lower level of education, legal or de facto spouses and mothers of young children. They were lower in individuals with a psychiatric disorder and in frequent and more recent users of mental health services according to the RAMQ's register. Conclusions: These findings support the hypotheses that social desirability and recall bias are likely to affect the self-reported use of mental health services in a population survey. They stress the need to refine the investigation of mental health services in population surveys and to combine survey and administrative data, whenever possible, to obtain an optimal estimation of the population need for mental health care.

Ellis, R. P., Fiebig, D. G., Johar, M., Jones, G., & Savage, E. (2013). Explaining Health Care Expenditure Variation: Large-Sample Evidence Using Linked Survey and Health Administrative Data. *Health Economics, 22*(9), 1093-1110. doi: 10.1002/hec.2916

Explaining individual, regional, and provider variation in health care spending is of enormous value to policymakers but is often hampered by the lack of individual level detail in universal public health systems because budgeted spending is often not attributable to specific individuals. Even rarer is self-reported survey information that helps explain this variation in large samples. In this paper, we link a cross-sectional survey of 267188 Australians age 45 and over to a panel dataset of annual healthcare costs calculated from several years of hospital, medical and pharmaceutical records. We use this data to distinguish between cost variations due to health shocks and those that are intrinsic (fixed) to an individual over three years. We find that high fixed expenditures are positively associated with age, especially older males, poor health, obesity, smoking, cancer, stroke and heart conditions. Being foreign born, speaking a foreign language at home and low income are more strongly associated with higher time-varying expenditures, suggesting greater exposure to adverse health shocks.

Emery, T. (2016). Administrative data linking: enriching administrative data with surveys. Paper presented at the European Conference on Quality in Offical Statistics, 31 May-3 June 2016, Madrid.

In social survey research there is a great deal of interest in enriching survey data with administrative data sources such as income, employment or even criminal records. However, the added value of such data linking to administrative data sources is rarely considered. Using data from the GGS in Sweden as an example, this paper outlines ways in which administrative data can be enriched by linking to a social survey. First, the survey process provides an opportunity to attain consent from respondents to link their data from administrative records (e.g. employment and birth records). In so doing, social surveys provide the key of consent for complex data linking and the subsequent ability to use administrative data to answer pressing social questions. Second, social surveys provide an opportunity to validate the data collection processes within administrative data collections. Third, the social surveys collect data themselves which is wholly absent from administrative records but is nonetheless of interest to both administrative data holders and social scientists. For example, the GGS contains data on the distribution of household work which, when taken in conjunction with administrative data provides key insights into gender roles throughout society. The analysis presented within the paper examines these three advantages and the degree to which they are evident in the case of the Swedish GGS. Given Sweden's strong administrative data tradition, it represents an example of how social survey data can supplement even a highly developed social statistics system.

European Union Agency for Fundamental Rights, & Council of Europe. (2014). *Handbook on European data protection law*. Retrieved from www.echr.coe.int/Documents/Handbook_data_protection_ENG.pdf

This handbook on European data protection law is jointly prepared by the European Union Agency for Fundamental Rights (FRA) and the Council of Europe together with the Registry of the European Court of Human Rights. The aim of this handbook is to raise awareness and improve knowledge of data protection rules in European Union and Council of Europe member states by serving as the main point of reference to which readers can turn. It is designed for non-specialist legal professionals, judges, national data protection authorities and other persons working in the field of data protection.

Gebhardt, H. P. (1990). The Legal Basis of Data Protection and Data Protection in the Field of Telecommunications in 7 Countries - Switzerland, France, the Federal-Republic-of-Germany, the United-Kingdom, Sweden, the United-States and Japan. *Telecommunication Journal, 57*(1), 37-44.

Groen, J. A. (2012). Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics, 28*(2), 173-198.

With increased attention to administrative data for statistical purposes, analyses of the quality of administrative data and comparisons to survey data are greatly needed. This article presents a methodology for identifying sources of error in administrative and survey data and for identifying sources of differences between administrative and survey estimates. The first part of the methodology is a statistical decomposition of the difference between administrative and survey estimates. The second part investigates the causes of measurement error and the factors associated with differences between what the same respondents report in the survey and the administrative records. I illustrate this methodology using a case study of the monthly employment figures gathered from U.S. businesses. This analysis demonstrates that both administrative data and survey data may contain errors and that reporting procedures are likely to differ between the two types of data. The article also identifies practical ways to assess data quality.

Groenewold, G., & Lessard-Phillips, L. (2012). Research methodology. In M. Crul, J. Schneider & F. Lelie (Eds.), The European Second Generation Compared. Does the Integration Context Matter? (pp. 39-56). Amsterdam: Amsterdam University Press.

This chapter discusses the TIES project (The Integration of the European Second Generation) research methodology. The following sections address the envisioned model sampling strategy, and sampling frame availability and constraints for each participating country and city. Apart from the problem of overcoming lack or deficiencies of nationally representative frames to sample from, the majority of the TIES survey target audience also proved difficult to contact and difficult to pin down for an interview. Low response rates raise doubts about whether responding persons can represent non-respondents in terms of personal characteristics and measured attitudes and opinions. For Amsterdam Rotterdam and Stockholm, this could be examined more closely because basic information on non-respondents was available in population registers. This information revealed that age, sex and marital status differences between responding and non-responding persons proved to be slight, lending support to claim statistical representativity of the collected data in these cities.

Groenewold, W. G. F., van Ginneken, J. K. S., & Masseria, C. (2008). Towards comparable statistics on mortality by socioeconomic status in EU member states. Methodology note. European Commission. Retrieved from http://ec.europa.eu/social/BlobServlet?docId=3958&langId=en

Overall levels of mortality have declined in all socioeconomic status (SES) groups in the EU, but there are indications that relative mortality differences between those in low and high SES

groups have increased. EU policy groups and departments expressed the need to address this issue but noted that comparable and high quality linked systems of data collection and statistics required for the monitoring and evaluation of policies addressing this issue, are not yet in place in all Member States. Although all National Statistical Institutes (NSI) in Member States collect and compile mortality data, they differ regarding the kind of SES characteristics that are collected. Educational attainment, occupation, and less so economic status, are main SES indicators for which data are collected and compiled. Yet evidence suggests that data and administrative data sources differ considerably concerning accessibility, completeness, coverage, quality, and employed database record-matching and -linkage methods. Furthermore, capabilities of NSIs differ as regards the routine derivation of mortality statistics by SES, and they often use different definitions and measures, notably of SES. Based on a rapid appraisal of data sources maintained by NSIs, a review of relevant studies on mortality differentialsby SES, including database record-linkage literature, our main recommendations are the following. First, Member States should preferably work towards improving and harmonizing their data sources on SES and mortality, by adopting a prospective, linked approach. Application of different record linkage methods of datasets should be explored, including deterministic and probabilistic methods and use of special-purpose software. Second, we recommend exploring other ways of deriving mortality statistics by SES, such as by (a) incorporating educational attainment and last occupation on death certificates; (b) covering the conventional SES indicators of educational attainment, occupational status and economic status in the forthcoming 2011 round of censuses in the EU; (c) collecting in the census the latter type of information, by proxy, of recently deceased household members; (d) covering the aforementioned issues in a sample survey built into the census (i.e. use of a more elaborate questionnaire in every k-th household), if costs of covering them in census questionnaires are too high. Third, we recommend that both absolute and relative measures are used in analyses and publications, such as rate ratios and rate differences in mortality of lowest versus highest socio-economic groups, relative index of inequality, and slope index of inequality. Fourth, to assist Members States, a technical cooperation and assistance project must be developed that starts out with theproduction of country-specific assessment studies covering: (a) organisational and technical details of existing health, mortality and SES data collection and compilation sources and procedures, including indicators and definitions used; (b) legislative, logistical, financial, technical, and human resource constraints, including training needs; (c) identification and formulation of feasible strategies to overcome these constraints. The evaluation of assessment reports at regional workshops should translate into the development and implementation of realistic work plans which, by the end of the project, result in (1) harmonized systems of data collection and compilation on mortality by SES, and (2) country reports, describing and analyzing health and mortality conditions by socioeconomic status, using comparable measures and methods of analysis, and common formats of presentation of findings

Haider, S. J., & Loughran, D. S. (2008). The effect of the social security earnings test on male labor supply - New evidence from survey and administrative data. *Journal of Human Resources, 43*(1), 57-87.

Despite numerous empirical studies, there is surprisingly little agreement about whether the Social Security earnings test affects male labor supply. In this paper, we provide a comprehensive analysis of the labor supply effects of the earnings test using longitudinal administrative earnings data and more commonly used survey data. We find that the response to the earnings test in survey data is obfuscated by measurement error and labor market rigidities. Accounting for these factors, our results suggest a consistent and substantial response to the earnings test, especially for younger men.

Hallinan, D., Frieclewald, M., & McCarthy, P. (2012). Citizens' perceptions of data protection and privacy in Europe. *Computer Law & Security Review, 28*(3), 263-272. doi: 10.1016/j.clsr.2012.03.005

Data protection and privacy gain social importance as technology and data flows play an ever greater role in shaping social structure. Despite this, understanding of public opinion on these issues is conspicuously lacking. This article is a meta-analysis of public opinion surveys on data protection and privacy focussed on EU citizens. The article firstly considers the understanding and awareness of the legal framework for protection as a solid manifestation of the complex concepts of data protection and privacy. This is followed by a consideration of perceptions of

privacy and data protection in relation to other social goals, focussing on the most visible of these contexts the debate surrounding privacy, data protection and security. The article then considers how citizens perceive the 'real world' environment in which data processing takes place, before finally considering the public's perception and evaluation of the operation of framework against environment.

Hill, K. (1987). Estimating census and death registration completeness. *Asian and Pacific Population Forum, 1*(3), 8-13.

Summary Death registration statistics, even when incomplete, can provide valuable information about mortality. In particular, the age structure of deaths can be used to estimate the completeness of registration, provided that this completeness does not vary substantially with age. Two methods of estimating the completeness of death registration from the distribution of deaths by age are described. The first is derived from stable population theory and requires an estimate of the rate of natural increase of the population, as well as assuming stability. However, the technique can also be used to generate simultaneously estimates of the rate of natural increase and of death registration completeness. The second method which requires two census age distributions and intercensal deaths by age, estimates the relative enumeration completeness of the two censuses as well as the completeness of death registration and requires only that the population be closed. Results are sensitive to overstatement of age. The methods are illustrated by being applied to figures from Thailand for the period 1960-70 and are found to work satisfactorily.

Hjollund, N. H., Larsen, F. B., & Andersen, J. H. (2007). Register-based follow-up of social benefits and other transfer payments: Accuracy and degree of completeness in a Danish interdepartmental administrative database compared with a population-based survey. *Scandinavian Journal of Public Health, 35*(5), 497-502. doi: 10.1080/14034940701271882

Background: Social consequences of disease may be subject to register based follow-up. A Danish database, DREAM, allows weekly follow-up of any public transfer payment. This study aimed to evaluate the feasibility of the register for use in public health research. Material and methods: The DREAM database includes information on all public transfer payments administered by Danish ministries, municipalities, and Statistics Denmark for all Danish citizens on a weekly basis since 1991. The DREAM database was compared with self-reported information on sources of income in a population survey from 2001 with about 5,000 participants. Results: According to DREAM, 80.2% of respondents had received some kind of transfer income since 1991. For the week they filled in the questionnaire, 9.0% had a record of labour-market-related benefit (unemployment benefit, social assistance, wage subsidy), 6.4% a health-related benefit (sickness benefit, vocational rehabilitation allowance, salary from subsidized jobs for persons with limited work capacity, anticipatory pension), 10.1% a voluntary retirement pension, while 74.4% had no record of transfer payment for that week. The predictive value of DREAM was 74.8% for health-related transfer payment and 98.2% for self-support. Among persons with a record of sickness benefit, 52.4% reported no transfer payment. Conclusion: The DREAM database is feasible for follow-up of social and economic consequences of disease. Respondents may be unaware of payments transferred by the public authorities to the employer, and in such cases DREAM may be the best source of information. The database is useful for public health research, but may also be useful for socioeconomic analyses of selection bias and dropout from other studies.

Hodge, J. G. (1999). The intersection of federal health information privacy and state administrative law: The protection of individual health data and workers' compensation. *Administrative Law Review, 51*(1), 117-144.

This article analyzes issues related to health information privacy in workers' compensation cases. Privacy of health information arises as an important issue in workers compensation cases because of the need to balance medical information as evidence with a workers' right to privacy. As such, the purpose, design and administration of workers' compensation cases as well as legal protections in federal and state constitutional, statutory and judicial law become important. Tension arises between federal interests in protecting health information privacy and state interests in regulating state workers' compensation cases free from federal intrusion.

Because information about health is so sensitive, there is a strong need to protect workers from mandatory disclosure of such information in order to bring a workers' compensation case. On the other hand, the need for states to be able to regulate workers' compensation and their need for such information is also valuable. After examining federal protection of health information collected, used and disclosed in state workers' compensation systems, the author proposes using state workers' compensation insurers and providers in the HHS proposed health insurance information privacy protections for proposed federal legislation. Although privacy interests are important in workers' compensation cases, in order for the system to succeed, individuals cannot ultimately control how their information is used. Society will benefit from a uniform and fair system to deal with employee injuries even if the disclosure of information in compulsory. The author advocates beginning by applying federal health information privacy systems to the systems already in place. The systems of health information privacy and workers' compensation can then be molded together to ensure the integrity of the information as well as the health of the workers.

Itoh, S., & Takano, M. (2011). A method to quantitatively assess confidentiality and potential usage of offical microdata in Japan. Paper presented at the Proceedings of the 58th World Statistical Congress, 2011, Session CPS020, Dublin.

This paper outlines disclosure avoidance methods that are currently used for official microdata in Japan, and examines micro-aggregation as a disclosure avoidance method. While perturbative methods such as additive noise and swapping including micro-aggregation are not currently adopted for official anonymized microdata in Japan, it is worth examining the applicability of perturbative methods. Second, this paper proposes methods of quantitatively assessing the usage potential and degree of confidentiality for official microdata, and conducts a comparative analysis of information loss and degree of confidentiality of masked data using the R-U map. This method enables a relative measurement of information loss and degree of confidentiality of masked data using perturbation for Japanese microdata.

Jasso, G., & Rosenzweig, M. R. (1982). Estimating the Emigration Rates of Legal Immigrants Using Administrative and Survey Data - the 1971 Cohort of Immigrants to the United-States. *Demography, 19*(3), 279-290. doi: Doi 10.2307/2060971

Based on administrative and survey data as well as data-based assumptions about the bounds on alien address reporting, this study provides estimates of the lower and upper bounds for the cumulative net emigration rates, by country and area of origin, of the FY1971 cohort of legal immigrants to the United States as of January 1979. The merged data indicate that the cumulative net emigration rate for the entire cohort could have been as high as 50 percent. Canadian emigration was probably between 51 and 55 percent. Emigration rates for legal immigrants from Central America, the Caribbean (excluding Cuba), and South America were at least as high as 50 percent, and could have been as high as 70 percent. Emigration rates for Koreans and Chinese could not have exceeded 22 percent over the same period.

Jenkins, S. P., Lynn, P., Jackle, A., & Sala, E. (2008). The Feasibility of Linking Household Survey and Administrative Record Data: New Evidence for Britain. *International Journal of Social Research Methodology, 11*(1), 29-43. doi: 10.1080/13645570701401602

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.

Kapteyn, A., & Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics, 25*(3), 513-551. doi: Doi 10.1086/513298

We provide both a theoretical and empirical analysis of the relation between administrative and survey data. By distinguishing between different sources of deviations between survey and administrative data we are able to reproduce several stylized facts. We illustrate the implications of different error sources for estimation in (simple) econometric models and find potentially very substantial biases. This article shows the sensitivity of some findings in the literature for the assumption that administrative data represent the truth. In particular, the common finding of substantial mean reversion in survey data largely goes away once we allow for a richer error structure.

Kennedy, E., & Millard, C. (2016). Data security and multi-factor authentication: Analysis of requirements under EU law and in selected EU Member States. Computer Law & Security Review, 32(1), 91-110.

Ensuring the security of personal data, particularly in terms of access controls, is becoming progressively more challenging. The most widely deployed authentication method, a user name plus a password, increasingly appears to be unfit-for-purpose. A more robust technique for maintaining the security of personal data is multi-factor authentication whereby two or more different types of credential are required. This approach is gaining traction, and in the European Union, some national data protection authorities are already recommending the use of multi-factor authentication as a means of complying with the obligation in the EU Data Protection Directive to take "appropriate technical and organisational measures to protect personal data". A proposal to replace that Directive with a General Data Protection Regulation is at an advanced stage in the EU legislative process with enhanced data security a central feature of the proposed reform. This article examines how the proposed Regulation would be likely to change the standard for data security both in general terms and in specific ways that might have an impact on the use of multi-factor authentication. Other sources of EU guidance are also considered, together with the position under the national laws and regulatory practices of six EU Member States.

Kim, C., & Tamborini, C. R. (2014). Response Error in Earnings: An Analysis of the Survey of Income and Program Participation Matched With Administrative Data. *Sociological Methods & Research, 43*(1), 39-72.

This article examines the problem of response error in survey earnings data. Comparing workers' earnings reports in the U.S. Census Bureau's Survey of Income and Program Participation (SIPP) to their detailed W-2 earnings records from the Social Security Administration, we employ ordinary least squares (OLS) and quantile regression models to assess the effects of earnings determinants and demographic variables on measurement errors in 2004 SIPP earnings in terms of bias and variance. Results show that measurement errors in earnings are not classical, but mean-reverting. The directions of bias for subpopulations are not constant, but varying across levels of earnings. Highly educated workers more correctly report their earnings than less educated workers at higher earnings levels, but they tend to over-report at lower earnings levels. Black workers with high earnings underreport to a greater degree than comparable whites, while black workers with low earnings over-report to a greater degree. Some subpopulations exhibit higher variances of measurement errors than others. Blacks, Hispanics, high school dropouts, part-year employed workers, and occupation switchers tend to misreport both over- and underreport their earnings rather than unilaterally in one direction. The implications of our findings are discussed.

Klassen, A. E., Lee, S. K., Barer, M., & Raina, P. (2005). Linking survey data with administrative health information: Characteristics associated with consent from a neonatal intensive care unit follow-up study. *Canadian Journal of Public Health-Revue Canadienne De Sante Publique, 96*(2), 151-154.

Background: Health services and population health research often depends on the ready availability of administrative health data. However, the linkage of survey-based data to administrative data for health research purposes has raised concerns about privacy. Our aim

was to compare consent rates to data linkage in two samples of caregivers and describe characteristics associated with consenters. Methods: Subjects included caregivers of children admitted at birth to neonatal intensive care units (NICU) in British Columbia and caregivers of a sample of healthy children. Caregivers were asked to sign a consent form enabling researchers to link the survey information with theirs and their child's provincially collected health records. Bivariate analysis identified sample characteristics associated with consent. These were entered into logistic regression models. Results: The sample included 1,140 of 2,221 NICU children and 393 of 718 healthy children. The overall response rate was 55% and the response rate for located families was 67.1%. Consent to data linkage with the child data was given by 71.6% of respondents and with caregiver data by 671% of respondents. Families of healthy children were as likely to provide consent as families of NICU children. Higher rates of consent were associated with being a biological parent, not requiring survey reminders, involvement in a parent support group, not working full-time, having less healthy children, multiple births and higher income. Conclusion: The level of consent achieved suggests that when given a choice, most people are willing to permit researcher access to their personal health information for research purposes. There is scope for educating the public about the nature and importance of research that combines survey and administrative data to address important health questions.

Korbmacher, J. M., & Schroeder, M. (2013). Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer. *Survey Research Methods, 7*(2), 115-131.

Linking survey data with administrative records is becoming more common in the social sciences in recent years. Regulatory frameworks require the respondent's consent to this procedure in most cases. Similar to non-response, non-consent may lead to selective samples and could pose a problem when using the combined data for analyses. Thus investigating the selectivity and the determinants of the consent decision is important in order to find ways to reduce non-consent. Adapting the survey participation model by Groves and Couper (1998), this paper identifies different areas influencing the respondents' willingness to consent. In addition to control variables at the individual and household level, two further areas of interest are included: the interview situation and the characteristics of the interviewer. A multilevel approach highlights the importance of the interviewer for the consent decision: the empty model shows an intra-class correlation of 55%, which can be reduced to 35% in a full model including interviewer variables. An additional analysis including measures of interviewer performance shows that there are further unobserved interviewer characteristics that influence the respondents consent decision. The results suggest that although respondent and household characteristics are important for the consent decision, a large part of the variation in the data is explained by the interviewers. This finding stresses the importance of the interviewers not only as an integral part in data collection efforts, but also as the direct link to gain a respondent's consent for linking survey data with administrative records.

Künn, S. (2015). The challenges of linking survey and administrative data IZA World of Labor 2015: 214. doi: 10.15185/izawol.214. http://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data.pdf

Linking survey and administrative data offers the potential for many new research opportunities for scientific and policy-related projects. While the number of linking projects in labor economics has been growing, the number is still very small. Growth in the number of projects has been slowed by concerns for individual privacy since consent to share administrative data is rarely available unless obtained explicitly through surveys that request such permission from respondents. As a result, legal constraints on the use of administrative data limit access to linked data and reduce the number of variables available for analysis because of the need to anonymize the data. Issues of privacy and consent remain the main challenge when linking both data sources. To ease these bottlenecks, policymakers, who have a lot to gain from the findings of research using linked data,should facilitate data linkage projects for scientific research. Doing so would result in the more efficient use of existing records and could also spark new research projects that may contribute novel insights and allow for drawing more reliable policy conclusions.

Lawrence, D., Christensen, D., Mitrou, F., Draper, G., Davis, G., McKeown, S., . . . Zubrick, S. R. (2012). Adjusting for under-identification of Aboriginal and/or Torres Strait Islander births in time series produced from birth records: Using record linkage of survey data and administrative data sources. *Bmc Medical Research Methodology, 12*.

Background: Statistical time series derived from administrative data sets form key indicators in measuring progress in addressing disadvantage in Aboriginal and Torres Strait Islander populations in Australia. However, inconsistencies in the reporting of Indigenous status can cause difficulties in producing reliable indicators. External data sources, such as survey data, provide a means of assessing the consistency of administrative data and may be used to adjust statistics based on administrative data sources. Methods: We used record linkage between a large-scale survey (the Western Australian Aboriginal Child Health Survey), and two administrative data sources (the Western Australia (WA) Register of Births and the WA Midwives' Notification System) to compare the degree of consistency in determining Indigenous status of children between the two sources. We then used a logistic regression model predicting probability of consistency between the two sources to estimate the probability of each record on the two administrative data sources being identified as being of Aboriginal and/or Torres Strait Islander origin in a survey. By summing these probabilities we produced model-adjusted time series of neonatal outcomes for Aboriginal and/or Torres Strait Islander births. Results: Compared to survey data, information based only on the two administrative data sources identified substantially fewer Aboriginal and/or Torres Strait Islander births. However, these births were not randomly distributed. Births of children identified as being of Aboriginal and/or Torres Strait Islander origin in the survey only were more likely to be living in urban areas, in less disadvantaged areas, and to have only one parent who identifies as being of Aboriginal and/or Torres Strait Islander origin, particularly the father. They were also more likely to have better health and wellbeing outcomes. Applying an adjustment model based on the linked survey data increased the estimated number of Aboriginal and/or Torres Strait Islander births in WA by around 25%, however this increase was accompanied by lower overall proportions of low birth weight and low gestational age babies. Conclusions: Record linkage of survey data to administrative data sets is useful to validate the quality of recording of demographic information in administrative data sources, and such information can be used to adjust for differential identification in administrative data.

Lifang G., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record Linkage: Current Practice and Future Directions. CMIS Technical Report No. 03/83. from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.8119

Record linkage is the task of quickly and accurately identifying records corresponding to the same entity from one or more data sources. Record linkage is also known as data cleaning, entity reconciliation or identification and the merge/purge problem. This paper presents the "standard" probabilistic record linkage model and the associated algorithm. Recent work in information retrieval, federated database systems and data mining have proposed alternatives to key components of the standard algorithm. The impact of these alternatives on the standard approach are assessed. The key question is whether and how these new alternatives are better in terms of time, accuracy and degree of automation for a particular record linkage application.

Lix, L. M., Yogendran, M. S., Shaw, S. Y., Targownick, L. E., Jones, J., & Bataineh, O. (2010). Comparing administrative and survey data for ascertaining cases of irritable bowel syndrome: a population-based investigation. *Bmc Health Services Research, 10*. doi: Artn 3110.1186/1472-6963-10-31

Background: Administrative and survey data are two key data sources for population-based health research about chronic disease. The objectives of this methodological paper are to: (1) estimate agreement between the two data sources for irritable bowel syndrome (IBS) and compare the results to those for inflammatory bowel disease (IBD); (2) compare the frequency of IBS-related diagnoses in administrative data for survey respondents with and without self-reported IBS, and (3) estimate IBS prevalence from both sources. Methods: This retrospective cohort study used linked administrative and health survey data for 5,134 adults from the province of Manitoba, Canada. Diagnoses in hospital and physician administrative data were investigated for respondents with self-reported IBS, IBD, and no bowel disorder. Agreement

between survey and administrative data was estimated using the kappa statistic. The chi(2) statistic tested the association between the frequency of IBS-related diagnoses and self-reported IBS. Crude, sex-specific, and age-specific IBS prevalence estimates were calculated from both sources. Results: Overall, 3.0% of the cohort had self-reported IBS, 0.8% had self-reported IBD, and 95.3% reported no bowel disorder. Agreement was poor to fair for IBS and substantially higher for IBD. The most frequent IBS-related diagnoses among the cohort were anxiety disorders (34.4%), symptoms of the abdomen and pelvis (26.9%), and diverticulitis of the intestine (10.6%). Crude IBS prevalence estimates from both sources were lower than those reported previously. Conclusions: Poor agreement between administrative and survey data for IBS may account for differences in the results of health services and outcomes research using these sources. Further research is needed to identify the optimal method(s) to ascertain IBS cases in both data sources.

Lujic, S., Watson, D. E., Randall, D. A., Simpson, J. M., & Jorm, L. R. (2014). Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia. *Bmj Open, 4*(9). doi: ARTN e005768 10.1136/bmjopen-2014-005768

Objectives: To investigate the nature and potential implications of under-reporting of morbidity information in administrative hospital data. Setting and participants: Retrospective analysis of linked self-report and administrative hospital data for 32 832 participants in the large-scale cohort study (45 and Up Study), who joined the study from 2006 to 2009 and who were admitted to 313 hospitals in New South Wales, Australia, for at least an overnight stay, up to a year prior to study entry. Outcome measures: Agreement between self-report and recording of six morbidities in administrative hospital data, and between-hospital variation and predictors of positive agreement between the two data sources. Results: Agreement between data sources was good for diabetes (kappa = 0.79); moderate for smoking (kappa = 0.59); fair for heart disease, stroke and hypertension (kappa = 0.40, kappa = 0.30 and kappa = 0.24, respectively); and poor for obesity (kappa = 0.09), indicating that a large number of individuals with self-reported morbidities did not have a corresponding diagnosis coded in their hospital records. Significant between-hospital variation was found (ranging from 8% of unexplained variation for diabetes to 22% for heart disease), with higher agreement in public and large hospitals, and hospitals with greater depth of coding. Conclusions: The recording of six common health conditions in administrative hospital data is highly variable, and for some conditions, very poor. To support more valid performance comparisons, it is important to stratify or control for factors that predict the completeness of recording, including hospital depth of coding and hospital type (public/private), and to increase efforts to standardise recording across hospitals. Studies using these conditions for risk adjustment should also be cautious of their use in smaller hospitals.

Luks, S., & Brady, H. E. (2003). Defining welfare spells - Coping with problems of survey responses and administrative data. *Evaluation Review, 27*(4), 395-420. doi: 10.1177/0193841X03254345

The authors explore how to define a welfare spell and how well surveys measure welfare spells. By comparing survey and administrative data from the Work Pays Demonstration Project in California on the receipt of Aid to Families with Dependent Children (AFDC), they find that a substantial amount of administrative churning occurs in administrative data. Through a mixing model of several break lengths, the authors find that a single definition of a break in welfare is not applicable to all respondents. Additionally, it appears that there is substantial variation in the break lengths respondents utilize. Finally, the authors show that the complexity of defining an accurate break in spells creates difficulties for detecting biases in survey responses.

Meints, M., Biermann, H., Bromba, M., Busch, C., Hornung, G., & Quiring-Kock, G. (2008). Biometric systems and data protection legislation in Germany. *2008 Fourth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Proceedings*, 1088-1093. doi: Doi 10.1109/Iih-Msp.2008.314

How can biometric systems be developed and run in compliance with European data protection legislation? The German TeleTrusT Association (www.teletrust.de) provides relevant information for manufacturers, vendors and users of biometric systems in a White Paper. The

following article gives an overview on most relevant data protection principles in the context of biometric systems, related threats and possible countermeasures, based on the mentioned White Paper.

Mellors, C., & Pollitt, D. (1984). Legislating for Privacy - Data Protection in Western-Europe. *Parliamentary Affairs, 37*(2), 199-215.

The article discusses about the importance of privacy and personal data and their protection in Western Europe. The author discusses the issues surrounding the protection of data. The pressure on governments of different countries in Western Europe into enacting the legislation to protect privacy and personal data is presented. He also mentions the issues encountered such as the definition of "privacy" and key features of legislation such as regulation and the revisions made by the British Parliament. An overview of other versions of data protection laws enacted in other countries in the western part of Europe is also presented.

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

This study expands our knowledge of consent in linking survey and administrative data by studying respondents' behaviour when consenting to link their own records and when consenting to link those of their children. It develops and tests a number of hypothesised mechanisms of consent, some of which were not explored in the past. The hypotheses cover: parental pride, privacy concerns, loyalty to the survey, pre-existing relations with the agency holding the data, and interviewer effects. The study uses data from the longitudinal Millennium Cohort Study to analyse the correlates of consent in multiple domains (i.e. linkage of education, health and economic records). The findings show that respondent's behaviour varies depending on the consent domain and on the person within the household for whom consent is sought. In particular, the cohort member's cognitive skills and the main respondent's privacy concerns have differential effects on consent. On the other hand, loyalty to the survey proxied by the longitudinal response history has a significant and strong impact on consent irrespective of the outcome. The findings also show that interviewers account for a large proportion of variations in consent even after controlling for the characteristics of the interviewer's assignment area. In total, it is possible to conclude that the significant impact of some of the correlates will lead to sample bias which needs to be accounted for when working with linked survey and administrative data.

Olesen, S. C., Butterworth, P., Jacomb, P., & Tait, R. J. (2012). Personal factors influence use of cervical cancer screening services: epidemiological survey and linked administrative data address the limitations of previous research. *Bmc Health Services Research, 12*.

Background: National screening programs have reduced cervical cancer mortality; however participation in these programs varies according to women's personal and social characteristics. Research into these inequalities has been limited by reliance on self-reported service use data that is potentially biased, or administrative data that lacks personal detail. We address these limitations and extend existing research by examining rates and correlates of cervical screening in a large epidemiological survey with linked administrative data. Methods: The cross-sectional sample included 1685 women aged 44-48 and 64-68 years from the Australian Capital Territory and Queanbeyan, Australia. Relative risk was assessed by logistic regression models and summary Population Attributable Risk (PAR) was used to quantify the effect of inequalities on rates of cervical cancer screening. Results: Overall, 60.5% of women participated in screening over the two-year period recommended by Australian guidelines. Screening participation was associated with having children, moderate or high use of health services, employment, reported lifetime history of drug use, and better physical functioning. Conversely, rates of cervical screening were lower amongst women who were older, reliant on welfare, obese, current smokers, reported childhood sexual abuse, and those with anxiety symptoms. A summary PAR showed that effective targeting of women with readily observable risk-factors (no children, no partner, receiving income support payments, not working, obese, current smoker, anxiety, poor physical health, and low overall health service use) could potentially reduce overall non-participation in screening by 74%. Conclusions: This study illustrates a valuable method for investigating the personal determinants of health service use

by combining representative survey data with linked administrative records. Reliable knowledge about the characteristics that predict uptake of cervical cancer screening services will inform targeted health promotion efforts.

Papakonstantinou, V. (2016). Exchange of Information and Data Protection in Cross-border Criminal Proceedings in Europe. *Common Market Law Review, 53*(2), 589-591.

In the past 10 years, the Member States of the European Union (EU) have intensified their exchange of information for the purposes of preventing and combating serious cross-border crime, as manifested in three main aspects. Firstly, there is a need to ensure the practical application of innovative principles (availability, mutual recognition) and concepts (Information Management Strategy, European Information Exchange Model) for tackling criminal organisations and networks that threaten the Internal Security of the EU. Secondly, there has been a gradual consolidation of EU agencies and bodies (Eurojust, Europol) aimed at promoting cooperation and dialogue among law enforcement officials and judicial authorities responsible for preventing and combating drug trafficking, trafficking in human beings, child pornography, and other serious trans-national offences. Thirdly, important EU information systems and databases (Prüm, SIS-II, ECRIS) have been created, enabling law enforcement and judicial authorities to gain access to essential information on criminal phenomena and organisations. Pursuing a practice-orientated approach, this work provides comprehensive coverage of all these measures, as well as the applicable rules governing data quality, data protection and data security. It is especially intended for law enforcement and judicial authorities who need to develop the appropriate expertise for the practical application of the above-mentioned principles. It also offers a solid basis of practical training material for police training centres and judicial schools.

Poulain, M., & Herm, A. (2013). Central Population Registers as a Source of Demographic Statistics in Europe. *Population, 68*(2), 215-247. doi: 10.3917/popu.1302.0215

Since their origins in seventeenth-century Sweden, population registers have been kept at local level, and computerization has now made it possible to establish national registers in most of the 30 European countries analysed in this article. As a result of these registers, the production of demographic statistics has entered a new era, with many advantages but also ethical controversies. New questions arise, such as the definition of residents, double counting and data confidentiality. This article describes and compares the operational principles of central registers in various EU countries, and how individual data are extracted in order to produce demographic statistics. It is now possible to regularly monitor the individual demographic trajectories of the entire population at national level, and to reveal interactions between the demographic behaviours of individuals in a single household. Given the many opportunities afforded by longitudinal analysis, support from researchers would be particularly beneficial, and efforts must be made to facilitate access to individual data.

Pratesi, M. (2016). *Analysis of Poverty Data by Small Area Estimation*: John Wiley & Sons.

A comprehensive guide to implementing SAE methods for poverty studies and poverty mapping. There is an increasingly urgent demand for poverty and living conditions data, in relation to local areas and/or subpopulations. Policy makers and stakeholders need indicators and maps of poverty and living conditions in order to formulate and implement policies, (re)distribute resources, and measure the effect of local policy actions. Small Area Estimation (SAE) plays a crucial role in producing statistically sound estimates for poverty mapping. This book offers a comprehensive source of information regarding the use of SAE methods adapted to these distinctive features of poverty data derived from surveys and administrative archives. The book covers the definition of poverty indicators, data collection, including surveys and administrative data sources, and methods to integrate them, the impact of sampling design, weighting and variance estimation, the issue of SAE modelling and robustness, the spatio-temporal modelling of poverty, and the SAE of the distribution function of income and inequalities. Examples of data analyses and applications are provided, and the book is

supported by a website describing scripts written in SAS or R software, which accompany the majority of the presented methods.

Preen, D. B., Holman, C. D. J., Lawrence, D. M., Baynham, N. J., & Semmens, J. B. (2004). Hospital chart review provided more accurate comorbidity information than data from a general practitioner survey or an administrative database. *Journal of Clinical Epidemiology, 57*(12), 1295-1304. doi: 10.1016/j.jclinepi.2004.03.016

Background and Objective: The accuracy of comorbidity data within the Western Australian Data Linkage System was evaluated by means of comparison with hospital charts and a general practitioner (GP) survey. Methods: Patients (n = 2,037) with a hospital admission from 1991 to 1996 were selected. Linked data were extracted for 100 comorbidities, categorized into 16 diagnostic chapters, for each hospital admission within a 5-year period. Clinical chart review and a GP survey were performed. Comorbidity occurrence in each data source and false-positive and false-negative diagnoses were ascertained. Results: Administrative data contained 45.5% of comorbidity recorded in hospital charts and under ascertained secondary conditions for all 16 diagnostic chapters. False-positive diagnoses were low for most conditions (range: 0-1.5%): however, a high occurrence of false negatives existed for all comorbidity chapters (range: 16.3-91.3%). GP-identified comorbidity was 20.0% greater than that found using administrative data but, with the exceptions of injury-poisoning and cutaneous-subcutaneous disease was less (42.0%) than that observed from hospital charts. Conclusion: Our results indicate that when accurate comorbidity data are crucial to health outcome research, hospital chart review (as opposed to using administrative data) may be required. Furthermore, surveying GPs. at least in Australia appears an unsatisfactory alternative to hospital charts for obtaining retrospective comorbidity information.

Rasner, A., Frick, J. R., & Grabka, M. M. (2013). Statistical Matching of Administrative and Survey Data: An Application to Wealth Inequality Analysis. *Sociological Methods & Research, 42*(2), 192-224.

Using population representative survey data from the German Socio-Economic Panel (SOEP) and administrative pension records from the Statutory Pension Insurance, the authors compare four statistical matching techniques to complement survey information on net worth with social security wealth (SSW) information from the administrative records. The unique properties of the linked data allow for a straight control of the quality of matches under each technique. Based on various evaluation criteria, Mahalanobis distance matching performs best. Exploiting the advantages of the newly assembled data, the authors include SSW in a wealth inequality analysis. Despite its quantitative relevance, SSW is thus far omitted from such analyses because adequate micro data are lacking. The inclusion of SSW doubles the level of net worth and decreases inequality by almost 25 percent. Moreover, the results reveal striking differences along occupational lines.

Sakshaug, J. W., & Kreuter, F. (2012). Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data. *Survey Research Methods, 6*(2), 113-122.

Administrative records are increasingly being linked to survey records to heighten the utility of the survey data. Respondent consent is usually needed to perform exact record linkage; however, not all respondents agree to this request and several studies have found significant differences between consenting and non-consenting respondents on the survey variables. To the extent that these survey variables are related to variables in the administrative data, the resulting administrative estimates can be biased due to non-consent. Estimating non-consent biases for linked administrative estimates is complicated by the fact that administrative records are typically not available for the non-consenting respondents. The present study can overcome this limitation by utilizing a unique data source, the German Panel Study "Labour Market and Social Security" (PASS), and linking the consent indicator to the administrative records (available for the entire sample). This situation permits the estimation of non-consent biases for administrative variables and avoids the need to link the survey responses. The impact of non-consent bias can be assessed relative to other sources of bias (nonresponse, measurement) for several administrative estimates. The results show that non-consent biases are present for few estimates, but are generally small relative to other sources of bias.

Sakshaug, J. W., Tutz, V., & Kreuter, F. (2013). Placement, Wording, and Interviewers: Identifying Correlates of Consent to Link Survey and Administrative Data. *Survey Research Methods, 7*(2), 133-144.

Record linkage is becoming more important as survey budgets are tightening while at the same time demands for more statistical information are rising. Not all respondents consent to linking their survey answers to administrative records, threatening inferences made from linked data sets. So far, several studies have identified respondent-level attributes that are correlated with the likelihood of providing consent (e.g., age, education), but these factors are outside the control of the survey designer. In the present study three factors that are under the control of the survey designer are evaluated to assess whether they impact respondents' likelihood of linkage consent: 1) the wording of the consent question; 2) the placement of the consent question and; 3) interviewer attributes (e.g., attitudes toward data sharing and consent, experience, expectations). Data from an experiment were used to assess the impact of the first two and data from an interviewer survey that was administered prior to the start of data collection are used to examine the third. The results show that in a telephone setting: 1) indicating time savings in the wording of the consent question had no effect on the consent rate; 2) placement of the consent question at the beginning of the questionnaire achieved a higher consent rate than at the end and; 3) interviewers' who themselves would be willing to consent to data linkage requests were more likely to obtain linkage consent from respondents.

Sandefur, J., & Glassman, A. (2015). The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics. *Journal of Development Studies, 51*(2), 116-132.

Across multiple African countries, discrepancies between administrative data and independent household surveys suggest official statistics systematically exaggerate development progress. We provide evidence for two distinct explanations of these discrepancies. First, governments misreport to foreign donors, as in the case of a results-based aid programme rewarding reported vaccination rates. Second, national governments are themselves misled by frontline service providers, as in the case of primary education, where official enrolment numbers diverged from survey estimates after funding shifted from user fees to per pupil government grants. Both syndromes highlight the need for incentive compatibility between data systems and funding rules.

Saydah, S. H., Geiss, L. S., Tierney, E., Benjamin, S. M., Engelgau, M., & Brancati, F. (2004). Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. *Annals of Epidemiology, 14*(7), 507-516. doi: 10.1016/j.annepidem.2003.09.016

PURPOSE: The ability to identify prevalent cases of diagnosed diabetes is crucial to monitoring preventative care practices and health outcomes among persons with diagnosed diabetes. METHODS: We conducted a comprehensive literature review to assess and summarize the validity of various strategies for identifying individuals with diagnosed diabetes and to examine the factors influencing the validity of these strategies. RESULTS: We found that studies using either administrative data or survey data were both adequately sensitive (i.e., identified the majority of cases of diagnosed diabetes) and highly specific (i.e., did not identify the individuals as having diabetes if they did not). In contrast, studies based on cause-of-death data from death certificates were not sensitive, failing to identify about 60% of decedents with diabetes and in most of these studies, researchers did not report specificity or positive predictive value. CONCLUSIONS: Surveillance is critical for tracking trends in diabetes and targeting diabetes prevention efforts. Several approaches can provide valuable data, although each has limitations. By understanding the limitations of the data, investigators will be able to estimate diabetes prevalence and improve surveillance of diabetes in the population. (C) 2004 Elsevier Inc. All rights reserved.

Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2015). Probabilistic record linkage. *International Journal of Epidemiology*. doi: 10.1093/ije/

Studies involving the use of probabilistic record linkage are becoming increasingly common. However, the methods underpinning probabilistic record linkage are not widely taught or

understood, and therefore these studies can appear to be a 'black box' research tool. In this article, we aim to describe the process of probabilistic record linkage through a simple exemplar. We first introduce the concept of deterministic linkage and contrast this with probabilistic linkage. We illustrate each step of the process using a simple exemplar and describe the data structure required to perform a probabilistic linkage. We describe the process of calculating and interpreting matched weights and how to convert matched weights into posterior probabilities of a match using Bayes theorem. We conclude this article with a brief discussion of some of the computational demands of record linkage, how you might assess the quality of your linkage algorithm, and how epidemiologists can maximize the value of their record-linked research using robust record linkage methods.

Schutze, B. (2013). Legal framework of data protection. Current requirements in Germany and requirements in planned European Union regulations. *Radiologe, 53*(5), 437-440. doi: 10.1007/s00117-013-2485-6

The federal system in Germany necessitates that in addition to federal laws, country and church-specific legislations must also be considered during the evaluation of relevant legal stipulations concerning data protection. Furthermore, there are also special legal regulations for hospitals in almost every federal state which are governed by the principle of subsidiarity: special legal regulations are to be preferentially used, so that findings from one federal state are difficult to transfer to another federal state. Patient data may only be used and processed without legal regulations with informed consent of the patient. The use of patient data for purposes of quality assurance, research and further education of students and doctors is possible under the present laws according to a positive weighting of interests. Patient data can also be exchanged via online services for the purposes of patient care; however, informed consent of the patient for medical online services is almost always unavoidable.

Sibley, L. M., Moineddin, R., Agha, M. M., & Glazier, R. H. (2010). Risk Adjustment Using Administrative Data-Based and Survey-Derived Methods for Explaining Physician Utilization. *Medical Care, 48*(2), 175-182.

Objectives: The objective of this study was to evaluate an administrative data-based risk adjustment method for predicting physician utilization and the contribution of survey-derived indicators of health status. The results of this study will support the use of administrative data for planning, reimbursement, and assessing equity of physician utilization. Methods: The Ontario portion of the 2000-2001 Canadian Community Health Survey was linked with administrative physician claims data from 2002-2003 and 2003-2004. Explanatory models of family physician (FP) and specialist physician (SP) utilization were run using demographic information and The Johns Hopkins University Adjusted Clinical Groups (ACG) Case-mix System. Survey-based measures of health status were then added to the models. The coefficient of determination, R, indicated the models' explanatory power. Results: The study sample consisted of 25,558 individuals aged 20 to 79 years representing approximately 7.8 million people. Over the 2 years of study period, 82.5% of the study population had a FP visit with a median of 6 visits and 53.2% had a SP visit with a median of I visit. The R 2 values based on administrative data alone were 33% and 21% for the frequency of FP and SP visits and 16% and 35% for having one or more visit to an FPs and SPs, respectively. The addition of the survey-based measures to the administrative data-based models produced less than a 2% increase in explanatory power for any outcome. Conclusion: Administrative data-based measures of morbidity burden are valid and useful indicators of future physician utilization. The survey-derived measures used in this study did not contribute significantly to models on the basis of administrative data-based measures. These findings support the future use of administrative data-based data and Adjusted Clinical Groups for planning, reimbursement, and research.

Stoddart, J., Chan, B., & Joly, Y. (2016). The European Union's Adequacy Approach to Privacy and International Data Sharing in Health Research. *Journal of Law Medicine & Ethics, 44*(1), 143-155.

The European Union (EU) approach to data protection consists of assessing the adequacy of the data protection offered by the laws of a particular jurisdiction against a set of principles that

includes purpose limitation, transparency, quality, proportionality, security, access, and rectification. The EU's Data Protection Directive sets conditions on the transfer of data to third countries by prohibiting Member States from transferring to such countries as have been deemed inadequate in terms of the data protection regimes. In theory, each jurisdiction is evaluated similarly and must be found fully compliant with the EU's data protection principles to be considered adequate. In practice, the inconsistency with which these evaluations are made presents a hurdle to international data-sharing and makes difficult the integration of different data-sharing approaches; in the 20 years since the Directive was first adopted, the laws of only five countries from outside of the EU, Economic Area, or the European Free Trade Agreement have been deemed adequate to engage in data transfers without the need for further administrative safeguards.

Sorensen, E. J. B. (2016). The post that wasn't: Facebook monitors everything users type and not publish. Computer Law & Security Review, 32(1), 146-151.

Revelations on the NSA surveillance programs have raised many questions in people's mind on the use of personal data and how personal data are analysed and for which purpose. But do users know that Facebook analyses everything they type and not publish? We (users) spend a lot of time considering what to post on Facebook. Should I argue that political point my friend has made? Should I comment on a friend's status? Do my friends need to see yet another picture of my pet or baby? Most of us have, at one point or another, started typing something and then regretted it and deleted the post. However, the code in our browser that powers Facebook remembers what we typed, even though we decide to delete it. This means that posts we explicitly choose not to share are not entirely private. Facebook calls these unpublished posts "self-censorship". A recent paper written by two Facebookers reveals how Facebook monitors and uses such unpublished thoughts. The question is whether this practice is covered by the Facebook Data Use Policy? The policy does not refer explicitly if Facebook manage self-censorship behavior. Moreover, while Facebook does not consider this practice a privacy violation, I argue that it clearly raises privacy and data protection issues. Under EU law, personal data can only be gathered legally under strict conditions, for a legitimate purpose. In this article I will analyse, using Facebook's self-censorship practices as an example, whether this practice is consistent with EU data protection law.

Tarozzi, A., & Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *Review of Economics and* Statistics, 91(4), 773-792. doi: 10.1162/rest.91.4.773

Recent years have seen widespread use of small-area maps based on census data enriched by relationships estimated from household surveys that predict variables, such as income, not covered by the census. The purpose is to obtain putatively precise estimates of poverty and inequality for small areas for which no or few observations are available in the survey. We argue that to usefully match survey and census data in this way requires a degree of spatial homogeneity for which the method provides no basis and which is unlikely to be satisfied in practice. We document the potential empirical relevance of such concerns using data from the 2000 census of Mexico.

Thompson, B. L., O'Connor, P., Boyle, R., Hindmarsh, M., Salem, N., Simmons, K. W., and Smith, S. M. (2001). Measuring clinical performance: Comparison and validity of telephone survey and administrative data. *Health Services Research, 36*(4), 813-825.

Objective. To compare and validate self-reported telephone survey and administrative data for two Health Plan Employer Data and Information Set (HEDIS) performance measures: mammography and diabetic retinal exams. Data. Sources/Study Setting. A telephone survey was administered to approximately 700 women and 600 persons with diabetes randomly chosen from each of two health maintenance organizations (HMOs). Study Design. Agreement of survey and administrative data was assessed by using kappa coefficients. Validity measures were assessed by comparing survey and administrative data results to a standard: when the two sources agreed, that was accepted as the standard; when they differed, confirmatory information was sought from medical records to establish the standard. When confirmatory information was not available ranges of estimates consistent with the data were constructed by first assuming that all persons for whom no information was available had received the service

and alternately that they had not received the service. Principal Findings. The kappas for mammography were .65 at both HMOs; for retinal exam they were .38 and .40. Sensitivity for both data sources was consistently high. However, specificity was lower for survey (range .44 to .66) than administrative data (.99 to 1.00). The positive predictive value was high for mammography using either data source but differed for retinal exam (survey .69 to .78; administrative data .99 to 1.00). Conclusions. Administrative and survey data performed consistently in both HMOs. Although administrative data appeared to have greater specificity than survey data the validity and utility of different data sources for performance measurement have only begun to be explored.

Trant, M., & Whitridge, P. (1998). Integration of administrative data with survey and census data. *Agricultural Statistics 2000, Proceedings*, 107-114.

Statistical systems are a consequence of evolution, and the level of data integration that is achieved is often an indication of the degree of system development. Almost all national statistical agencies in the world integrate administrative data with their survey and census information to some degree in order to complement, supplement or replace survey information, or to assist with frame maintenance. This paper reconstructs the development and evolution of the Canadian agricultural statistical system as it relates to the expanding and increasingly important role of administrative data. The degree to which administrative data are integrated depends on a number of factors, the most important being: (1) the degree of maturity of the country's statistical system, (2) the quality and the amount of information available from the government's administrative and regulatory programs, (3) well-trained and experienced staff, (4) funding, and (5) cooperation among government agencies. Most countries appear to have gone through an evolutionary process in establishing their agricultural statistical system and most systems are developed with an internal capacity for renewal and adjustment. This allows them to respond to changing conditions and needs, and to remain relevant. The actual route that is followed, however, is highly dependent on the amount of resources available for the program, the availability of experienced professionals to develop and maintain the system, and the statistical toolbox that they are able to use.

Townend, D. *(2016). EU Laws on Privacy in Genomic Databases and Biobanking. Journal of Law Medicine & Ethics, 44(1), 128-142.*

Both the European Union and the Council of Europe have a bearing on privacy in genomic databases and biobanking. In terms of legislation, the processing of personal data as it relates to the right to privacy is currently largely regulated in Europe by Directive 95/46/EC, which requires that processing be fair and lawful and follow a set of principles, meaning that the data be processed only for stated purposes, be sufficient for the purposes of the processing, be kept only for so long as is necessary to achieve those purposes, and be kept securely and only in an identifiable state for such time as is necessary for the processing. The European privacy regime does not require the de-identification (anonymization) of personal data used in genomic databases or biobanks, and alongside this practice informed consent as well as governance and oversight mechanisms provide for the protection of genomic data.

United, N. (1983). Manual X: Indirect Techniques for Demographic Estimation. New York: United Nations.

Direct and indirect methods of estimating levels and age-patterns in mortality and fertility, using survey, census, vital statistics and population register data. Estimates based on these different types of data sources are compared to arrive at final estimates of fertility and mortality.

Van den Akker, M., Buntinx, F., Metsemakers, J. F. M., & Knottnerus, J. A. (1998). Morbidity in responders and non-responders in a register-based population survey. *Family Practice, 15*(3), 261-263. doi: DOI 10.1093/fampra/15.3.261

Background. Non-response analysis is often restricted to the influence of age, sex and socioeconomic status on response status. In this study the health status of responders and non-responders was also compared. Results. Responders were comparable to non-responders with regard to the number of diagnosed disorders as well as to the prevalences of disorders within body systems. Non-responders only showed psychological disorders more often.

Conclusion. It is useful to assess the relation between non-response and morbidity patterns in other studies as well, in order to detect selective non-response and bias.

Van der Molen, I. N., & Commers, M. J. (2013). Unresolved legal questions in cross-border health care in Europe: liability and data protection. *Public Health, 127*(11), 987-993. doi: 10.1016/j.puhe.2013.08.020

Objectives: Directive 2011/24/EU was designed to clarify the rights of EU citizens in evaluating, accessing and obtaining reimbursement for cross-border care. Based on three regional case studies, the authors attempted to assess the added value of the Directive in helping clarify issues in to two key areas that have been identified as barriers to cross-border care: liability and data protection. Study design: Qualitative case study employing secondary data sources including research of jurisprudence that set up a Legal framework as a base to investigate liability and data protection in the context of cross-border projects. Methods: By means of three case studies that have tackled liability and data protection hurdles in cross-border care implementation, this article attempts to provide insight into legal certainty and uncertainty regarding cross-border care in Europe. Results: The case studies reveal that the Directive has not resolved core uncertainties related to liability and data protection issues within cross-border health care. Some issues related to the practice of cross-border health care in Europe have been further clarified by the Directive and some direction has been given to possible solutions for issues connected to liability and data protection. Conclusions: Directive 2011/24/EU is clearly a transposition of existing regulations on data protection and ECJ case law, plus a set of additional, mostly, voluntary rules that might enhance regional border cooperation. Therefore, as shown in the case studies, a practical and case by case approach is still necessary in designing and providing cross-border care.

Warburton, R. N., & Warburton, W. P. (2004). Canada needs better data for evidence-based policy: Inconsistencies between administrative and survey data on welfare dependence and education. *Canadian Public Policy-Analyse De Politiques, 30*(3), 241-255. doi: Doi 10.2307/3552301

This study compares administrative and survey data on BC welfare (social assistance) recipients, to test whether survey data is sufficiently accurate for use in policy-oriented research. BC welfare and education data is compared to the 1994 Public Use Microdata (BC sample) of Statistics Canada's Survey of Labour and Income Dynamics (SLID). BC 1994 SLID significantly understates welfare dependence, and overstates education levels of BC welfare recipients. Statistics Canada should lead a national initiative to make provincial administrative datasets available for research; and should use these data to improve key national longitudinal social research surveys such as SLID, NLSCY, and NPHS.

Zaslavsky, A. M., Schenker, N., & Belin, T. R. (2001). Downweighting influential clusters in surveys: Application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association, 96*(455), 858-869. doi: Doi 10.1198/016214501753208889

Certain clusters may be extremely influential on survey estimates and consequently contribute disproportionately to their variance, We propose a general approach to estimation that downweighs highly influential clusters, with the amount of downweighting based on M-estimation applied to the empirical influence of the clusters, The method is motivated by a problem in census coverage estimation, and we illustrate it by using data from the 1990 Post Enumeration Survey (PES). In this context, an objective, prespecified methodology for handling influential observations is essential to avoid having to justify judgmental post hoc adjustment of weights. In 1990, both extreme weights and large errors in the census led to extreme influence. We estimated influence by Taylor linearization of the survey estimator, and we applied M-estimators based on the t distribution and the Huber psi -function, As predicted by theory, the robust procedures greatly reduced the estimated variance of estimated coverage rates, more so than did truncation of weights, On the other hand, the procedure may introduce bias into survey estimates when the distributions of the influence statistics are asymmetric. We consider the properties of the estimators in the presence of asymmetry, and we demonstrate techniques for assessing the bias-variance trade-off, finding that estimated mean squared error is reduced

by applying the robust procedure to our dataset. We also suggest PES design improvements to reduce the impact of influential clusters.

Zaslavsky, A. M., & Wolfgang, G. S. (1993). Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative-List Data. *Journal of Business & Economic Statistics, 11*(3), 279-288.

Dual-system measurement of census coverage using a post-enumeration survey has been criticized for correlation bias, resulting when responses to the census and survey are not independent. A third system provides additional information to assess that independence. This study focuses on urban Black male adults, using data from the 1988 Dress Rehearsal Census and its Post-Enumeration Survey and from other government sources. Results using a variety of models confirm that their population is underestimated by dual-system methods. Problems involving classification and matching errors are also discussed. The results suggest that triple-system modelling has great potential for more precise estimation of the hard-to-count population and its census coverage.

Zika, E., den Baumen, T. S. I., Kaye, J., Brand, A., & Ibarreta, D. (2008). Sample, data use and protection in biobanking in Europe: legal issues. *Pharmacogenomics, 9*(6), 773-781. doi: 10.2217/14622416.9.6.773

The sharing of samples and data stored in biobanks for research has implications for donor privacy, but also raises questions on the regulation of research within Europe. Many legal documents and principles within Europe, with a direct impact on biobanking, have not been developed specifically to support this activity. Moreover, while some new regulations have been set up at national level, there are many variations in the definitions, scope and purpose of these legal instruments. This has resulted in unnecessary hurdles for genome-based research, particularly if samples are shared across national borders. The question is also raised on whether new, specific legislative and governance frameworks designed for biobanking are needed, or whether it is sufficient to modify current general law and to develop specific guidelines, or to accommodate issues raised by biobanking in the current regulation. A workshop with experts from academia and industry, lawyers, national data protection authorities, representatives from the European Commission and the European Data Protection Supervisor was held to review the existing legal bottlenecks and future needs of biobanking, with special regard to the collection, exchange and linkage of samples and data. This report presents highlights of the presentations and discussions from the workshop held in Sevilla, Spain, in March 2007 and the conclusions that followed. The workshop focused on the internal linkage of data and samples stored in a biobank, and the external linkage of biobanks with secondary information resources, such as cancer registries.