Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2016: 1–13

DOI: 10.1177/2053951716684143
journals.sagepub.com/home/bds
SAGE

Administrative social science data: The challenge of reproducible research

Christopher J Playford[1], Vernon Gayle[1], Roxanne Connelly[2] and Alasdair JG Gray[3]

Abstract

Powerful new social science data resources are emerging. One particularly important source is administrative data, which were originally collected for organisational purposes but often contain information that is suitable for social science research. In this paper we outline the concept of reproducible research in relation to micro-level administrative social science data. Our central claim is that a planned and organised workflow is essential for high quality research using micro-level administrative social science data. We argue that it is essential for researchers to share research code, because code sharing enables the elements of reproducible research. First, it enables results to be duplicated and therefore allows the accuracy and validity of analyses to be evaluated. Second, it facilitates further tests of the robustness of the original piece of research. Drawing on insights from computer science and other disciplines that have been engaged in e-Research we discuss and advocate the use of Git repositories to provide a useable and effective solution to research code sharing and rendering social science research using micro-level administrative data reproducible.

Keywords

Big Data, administrative data, reproducibility, replication, workflow, Git

Introduction

The known universe of data that are available to social science researchers is ever expanding, and the second decade of the 21st Century is characterised by the explosion of new forms of data. The increased processing speed of computers and the expansion of affordable storage capacity present exciting opportunities for social science research. The result is that empirical studies in social science disciplines such as sociology are likely to become increasingly computationally intensive. Because of these rapid changes in both the data and the computational landscape we conjecture that social scientists need to re-think aspects of the research process.

King (2011) asserted that there are large challenges associated with using new forms of social science data (especially with accessing, analysing, preserving, and protecting information). In this paper we address some of the challenges associated with undertaking reproducible social science research with these new forms of data. There are a wide variety of data types and analytical techniques used within and across the disciplines and sub-disciplines that constitute the social sciences. In this paper we concentrate on the statistical analysis of large-scale and complex data sets which contain information on individuals. An exciting and emerging source of large-scale social science data is administrative data where information has originally been collected to organise, manage, monitor or deliver services but these data have measures that are suitable for social research (Woollard, 2014). Undertaking

[1]Administrative Data Research Centre – Scotland, School of Social and Political Science, University of Edinburgh, UK
[2]Department of Sociology, University of Warwick, UK
[3]School of Mathematical & Computer Sciences, Heriot-Watt University, UK

Corresponding author:
Christopher J Playford, Administrative Data Research Centre – Scotland, University of Edinburgh, Bioquarter Nine, 9 Little France Road, Edinburgh EH16 4UX, UK.
Email: chris.playford@ed.ac.uk

reproducible research using administrative social science data is the overall theme of this paper.

The more general issue of reproducibility in research is pithily summarized by the Yale Law School Roundtable on Data and Code Sharing (2010: 8) who conclude that:

> 'Computation is becoming central to the scientific enterprise, but the prevalence of relaxed attitudes about communicating computational experiments' details and the validation of results is causing a large and growing credibility gap. Generating verifiable knowledge has long been scientific discovery's central goal, yet today it's impossible to verify most of the computational results that scientists present at conferences and in papers.'

There is a more general call for extra materials that enable researchers to understand, evaluate and build upon prior work to be routinely provided alongside research publications. These materials should include sufficient information for a third party to reproduce results without any additional information from the authors (Diggle, 2015; King, 1995, 2003). High quality academic journals such as *Science*, *American Economic Review*, *Econometrica* and the *Review of Economic Studies* now require supporting computer code that is involved in the creation and analysis of data (Hanson et al., 2011; McCullough et al., 2008). Over 500 journals across all disciplines are now signatories of the *Transparency and Openness Promotion (TOP) Guidelines*, which require data and code sharing standards.[1] These guidelines provide details on transparency of data sharing, analytical methods, research materials, design, preregistration of studies and plans, replication and citation standards.[2] Wider discussions of the issue of research reproducibility are also currently taking place around the world, for example the Berkeley Initiative for Transparency in the Social Sciences[3] and Open Science Collaboration (2015).

Concern about the lack of reproducibility of research persists among scientists across a range of academic disciplines ('Reality check on reproducibility' [Editorial], 2016). Jahnke et al. (2012) identify a series of problems relating to data management and curation practices among university researchers including a lack of formal training, a lack of concern for the long-term preservation of data, the demands of publication outweighing good practices, documentation only being of interest if it directly assists the researcher, and a lack of effective collaboration tools.

The focus of this paper are the challenges encountered in undertaking 'reproducible research' (i.e. research which can be 'consistently repeated') using large-scale administrative social science data. Sharing research code is not currently a widespread practice within the social sciences, and is especially rare in disciplines such as sociology and social geography. We will argue that sharing research code is central and critical for achieving reproducibility. There are useful insights from current practices that are common in areas such as computer science and e-research that are germane to improving reproducibility in social science research.

What is reproducibility?

There are some differences in the definition of reproducibility across the social sciences and other academic areas engaged in e-Research. We use the terminology 'reproducibility' to describe the practice of producing social research which can be 'consistently repeated'. Following Janz (2015) we divide reproducibility into two related stages. The first stage is 'duplication'. A study can be duplicated if information is made available which ensures that consistent results can be produced using the same data and applying the same analytical techniques. Janz (2015) uses the term 'replication' to refer to the next stage. A replication study can 'duplicate' the original findings but also further tests the robustness of the original piece of research, for example by employing new or additional measures, data or methods.

We consider that there are four pillars of wisdom that inform successful statistical social science data analyses. The four pillars of wisdom are accuracy, efficiency, transparency and reproducibility. Accuracy relates to minimising information loss and errors in data construction, data analysis and research outputs. Efficiency relates to maximising the features offered by software, and when possible automating actions. Transparency is central to good social science data analysis practices. When work is appropriately transparent questions of the 'who, what, where, when and why' variety are easily answered. Reproducibility is central to good social science. In essence work is reproducible when it can first be 'duplicated' and then 'replicated'.

Successful research outcomes are far more likely if the analysis of large-scale social science data sets is guided by a planned workflow (Long, 2009). The workflow refers to a coordinated framework for conducting social science data analyses. The workflow includes planning, organising, executing and documenting analyses. The initial steps are likely to include bureaucratic activities such as applying for ethical approval, applying for access to the data, and gaining access to the data. This is likely to be followed by computational activities which begin with enabling data for analyses. The later steps are likely to include analysing data, presenting results, refining results, writing up and then publishing findings. The final steps will include

archiving files of data and results, and then rendering them 'reproducible'.

Recognised good practices in orchestrating a suitable workflow are as important in the analysis of administrative data as they are in the analysis of other large-scale social science data sets (e.g. social surveys). Long (2009) has provided an extensive, and almost rabbinical, account of good practices that we suggest analysts of micro-level administrative social science data should consult. In the passage below, we précis this work and make relevant connections with the emerging practice of administrative social science data analysis.

Central to the workflow is the concept of having an 'audit trail'. The 'audit trail' is nothing more than a chronological account of the activities undertaken in the data analytical process. An alternative visualisation is that it is the breadcrumb trail of the research process. The audit trail is important because within the statistical analysis of social science data sets minor decisions have major consequences. Keeping track of even the most seemingly minor actions in the workflow is therefore important as it facilitates transparency, and makes contributions to efficiency and accuracy and ultimately to the overall success of the research project (this is also advocated by Sandve et al., 2013). Long (2009: 296) argues that the '*the provenance of every result should be documented*.' Working towards this aim we believe that it is imperative for data analysts working with micro-level administrative social science data to create and record formal plans.

Statistical analysis using micro-level administrative social science data

Among the numerous 'Big Data' resources offering scope for social science research, a particularly valuable source is administrative data. A key feature of administrative data is that they were not originally collected for research purposes (Connelly et al., 2016). Administrative social science data may also be large, complex and multi-dimensional. Historically, social scientists have had very limited access to administrative records, with the exception of the register-based data sets of the Nordic counties (Figlio et al., 2015; Wallgren and Wallgren, 2007). The state of access to administrative data for social science research is at varying stages in the USA (Card et al., 2010), Canada (Doiron et al., 2013) and Western Australia (Holman et al., 2008). In the UK, the Economic and Social Research Council (ESRC) has recently funded the Administrative Data Research Network (ADRN)[4] which aims to appropriately open up access to a plethora of data that have been recorded in databases and files in various government departments which researchers have previously found hard to gain access to.

The overall goal of the UK ADRN is to provide social researchers with access to linked individual-level data from Government Departments and other agencies that routinely collect data relevant to social and economic research.[5] The ADRN will allow researchers to gain carefully supervised access to data to undertake studies that are ethical and feasible.[6] Individuals contained within micro-level administrative social science data sets are potentially identifiable, so the ADRN removes personal identifiers from the records accessible to researchers (for further details, see Dibben et al., 2015). The bar for gaining access to administrative data in the UK is set high because a great deal of work is required to link data and to get de-identified data ready for researchers to analyse. The outcome will be valuable new sources of social science data. These data will support detailed empirical analyses of social and economic life in contemporary Britain.

We use the term 'micro-level administrative social science data' to describe the wealth of new research data resources about individuals that are emerging as a result of recent efforts to make administrative data available to social science researchers. We use the prefix 'micro-level' because these data are non-aggregate and have a resolution that is suited to the analysis of individuals, couples, families and households, and because they are appropriate for investigating micro-level social processes. This contrasts with macro-level data where the units are aggregated, for example regions or nation states.[7] The forms of micro-level administrative data that we focus on here are suitable for social science research and similar in structure to other widely used social science data resources such as large-scale social surveys.

Much of the time spent by social scientists analysing micro-level administrative social science data will be analysing the familiar rectangular variable by case matrix, where a variable is recorded in each column and each case is allocated to a row.[8] When micro-level administrative data are organised for conventional social science research (e.g. the application of multivariate techniques such as statistical models) they are indistinguishable from variable by case matrices that are produced from data collected by large-scale social surveys. A characteristic of these micro-level administrative social science data sets is that they usually have a large number of observations (n), for example individuals, but a smaller number of social science related explanatory variables (k) than would be the case for social surveys.

There is absolutely nothing that convinces us that when analysing a variable by case matrix of micro-level administrative social science data, we can ignore the helpful lessons that have emerged from many decades of research in statistics and statistically orientated

areas of social science (and particularly the specialist area of econometrics). For example if a micro-level administrative social science data set has repeated measurements on the same individuals the usual problems associated with non-independence of observations, or the possibility of residual heterogeneity will not evaporate simply because the data are from an administrative source rather than a social survey. We argue that it is preferable to build on existing methodological traditions of social science data analysis than to overlook these approaches (Connelly et al., 2016; Harford, 2014). Therefore we see the analysis of micro-level administrative social science data as a special case of the more general activity of undertaking statistical analyses of large-scale social science data sets. There are a series of practical methodological issues which are specific to micro-level administrative social science data sets, but we argue that these issues are best understood by drawing on the existing methodological knowledge base in statistics and social science, and by drawing on insights from computer science and other disciplines that have been engaged in e-Research[9] (Hey et al., 2009).

The importance of using research code

In this section we advocate the sharing of research code and point to a series of related 'good' practices that will improve the reproducibility of administrative social science data analysis. Whilst the practices we advocate may appear to be routine to computer programmers, very few social scientists have formal training in computer science (see also Gentzkow and Shapiro, 2014). Our strong advice is that researchers using micro-level administrative social science data must never undertake manual manipulations of data that are undocumented (this point is emphasised by Sandve et al., 2013). Concurrently, we strongly warn against undertaking data analyses using Graphical User Interfaces (e.g. point and click methods) or using their software in an interactive model because these approaches are usually undocumented. We categorically state that it is imperative for researchers using micro-level administrative social science data to undertake their data analyses using syntax files. The use of syntax files is of vital importance to ensure that analyses are reproducible and transparent to others (Long and Freese, 2014; Treiman, 2009). Syntax files are text documents that contain code that issues the commands to statistical software (see Boslaugh, 2005). The term 'syntax file' was popularised in the social sciences by the programme SPSS, however we use the term to refer to all command files in statistical data analysis software packages, such as do files in Stata. We also consider R scripts to be syntax files because similarly they command data analyses. In this paper we also use the terms 'research code' and syntax files interchangeably.[10] Syntax files should be robust and be able to produce exactly the same result each time they are executed. Syntax files should be legible and be well-annotated with commentary so that it is easy for the reader to understand exactly what is being undertaken (Long, 2009: 51). Attempting to reproduce analyses, for example in response to requests from reviewers, without recourse to the syntax that produced both the data and the analyses is extremely time consuming and highly prone to errors (Freese, 2007). Use of syntax files enhances the accuracy, efficiency, transparency and reproducibility of research.

Challenges of reproducibility with micro-level administrative social science data

In this section we identify salient challenges to undertaking reproducible research with micro-level administrative social science data. These challenges include data access, data retention, working with dynamic data and difficulties in undertaking exploratory data analysis.

In addition to all the usual challenges that a social scientist will face when analysing a large-scale social science data set, micro-level administrative social science data sets often have heavy restrictions placed upon their access and use. This will frequently mean that the data analyst does not have desktop access to the data set. Many micro-level administrative social science data sets can only be analysed within secure environments. These are approved locations with special security arrangements. Secure environments are usually located at National Statistical Agencies, within some government departments, and at some universities with officially sanctioned facilities.[11] Researchers usually have to book time at these facilities, will have to travel to the secure environment, and access within these special settings will be supervised. To enhance security, researchers do not have access to the Internet when working in secure environments and any files that a researcher might wish to bring in to the environment are checked by an official member of staff. Researchers will want to use their time within secure environments as effectively as possible and having a planned workflow is therefore critical.

Some micro-level administrative social science data sets are not accessed directly. For example when analysing the Canadian Survey of Labour and Income Dynamics, which contains administrative data, users write programmes to send to Statistics Canada in electronic form which are then run by Statistics Canada staff on the data set (see Giles, 2001: 374–375). The output is then reviewed by the staff to ensure that no

risk to privacy exists, and then the results are delivered to the data analyst.[12] This means that the data analyst will not be able to use their statistical data analysis software in an interactive mode (e.g. via a graphical user interface). A planned workflow is therefore critical because a third party is involved in executing data analysis operations.

A critical feature of social science research using micro-level administrative data is that people should not be recognised or linked in a way that could infringe on privacy, in much the same way as they would be anonymised in a social survey data set. Therefore micro-level administrative social science data often have special constraints placed on their availability. The TOP guidelines acknowledge that there are exceptions where it may not be possible to make data publicly available. In this eventuality they advise that contributors should explain these restrictions to their audience, describe access procedures, provide details of software and documentation used, and provide access to data and material where restrictions do not apply.[13] Our position is that through better use of meta-information (whether provided by the data provider or created during the processes of extracting, linking and analysing the data), and through sharing the code that has been used in both data creation and analyses, substantial progress can be made towards the challenge of making the analysis of micro-level administrative social science data more reproducible.

Data retention policies are a further challenge when working with administrative social science data. For example in the UK, the ADRN Data Retention and Destruction Policy[14] specifies that research data will be archived for a maximum of five years. The Organisation for Economic Co-operation and Development (OECD) suggest that research funding agencies and research institutions should consider the long-term retention of data when evaluating projects, in order to deliver sustained public benefits (OECD, 2007). We advance the obvious argument that the preservation of research data are essential for reproducible research and strongly advocate that such policies be reviewed to ensure that data are retained in perpetuity along with the research code that has been used.

The micro-level social science data which emerge from administrative settings are sometimes the product of dynamic systems. These are systems which organise, manage, deliver and monitor services. They are typically recorded in databases and the contents can potentially change over time. Therefore there is a requirement to accurately document the exact data which are included in a social science analysis. This can typically be achieved by recording the code used to extract the data and the date and time parameters of the original data extract from the dynamic system.

A substantial activity in the analysis of any large-scale social science data set will be exploratory data analyses (see Marsh and Elliott, 2008; Tukey, 1977). This can be considered as the initial stage of data analysis, where researchers begin to understand the main characteristics of the data set, explore ideas and make initial inquiries. In more conventional analyses, for example using a household survey, a researcher can begin exploring the data set as soon as they have gained access. In many cases special arrangements will have been developed for researchers to explore a survey data set prior to gaining access to it. Notable examples include The (German) Socio-Economic Panel[15] and the British Household Panel Survey.[16] Exploratory data analysis of some social surveys has been supported through the development of NESSTAR,[17] a software system for data publishing and exploration. NESSTAR enables data providers to disseminate their data on the web so that users can search, browse and undertake exploratory data analysis online. Some micro-level administrative social science data sets do provide detailed information on their content, for example the Scottish Longitudinal Study has an online data dictionary[18] and the National Pupil Database provides detailed online materials.[19] At the current time there are few facilities to search micro-level administrative data resources[20] and to easily undertake the necessary exploratory data analyses that form a routine and required part of the process of social science research.

The case for research code sharing

Large-scale social science data sets, for example national surveys, are routinely made available via national data archives and repositories.[21] These data sets are provided to social science data analysts in a general format that can support a wide spectrum of potential analyses. It is typical for centres providing access to micro-level administrative social science data to provide data analysis software which is programmable using syntax. For example, the ADRN provide data sets for a range of software including SPSS, Stata and SAS files. We use the term 'data enabling' to describe the stage between downloading the social science data set and beginning to undertake statistical analyses. 'Data enabling' comprises tasks associated with preparing and enhancing data for statistical analysis, such as recoding measures, constructing new variables and linking data sets (Blum et al., 2009; Lambert and Gayle, 2008). 'Data enabling' is a substantial part of the research process and its importance is often overlooked. The time required to 'enable data' is frequently underestimated, even by more experienced social science data analysts. A planned workflow is critical for accurate, efficient, transparent and reproducible 'data enabling'.

Large-scale survey data sets typically undergo extensive amounts of data preparation, for example cleaning, cross-checking, testing and validating, before they are released for social science data analysis. This work is usually undertaken by the survey data collection agency or by the national data archive or data provision service. We refer to this process as 'data pre-enabling'. Micro-level administrative data have not primarily been collected for social science data analysis, and data sets will frequently be the integration of multiple data resources (Connelly et al., 2016). Working with micro-level administrative data will typically involve joining together and restructuring administrative data sets into a suitable format with all the pieces of information required to answer a social science research question (see Elias, 2014). In the UK for example, some 'data pre-enabling' tasks will have also been carried out by the data provider and by Trusted Third Parties[22] (TTPs) in the form of code used to extract and link the data sets. In contrast to working with social surveys, researchers working with micro-level administrative social science data may have to undertake both 'data pre-enabling' and 'data enabling' tasks before they can begin statistical analyses. In these circumstances a planned workflow is critical for both 'data pre-enabling' and 'data enabling' to be accurate, efficient, transparent and reproducible.

Our argument is that syntax files developed as part of the workflow in the process of both 'data pre-enabling' and 'data enabling' should more routinely be shared. The re-use and modification of existing coding within syntax files offers the potential to make an overall contribution to more accurate and efficient administrative social science data analyses that are transparent and reproducible. We argue that this should also include the code used by data providers and TTPs to extract and link the micro-level administrative social science data. Retaining and sharing code which has been suitably cleared through the normal protocol of statistical disclosure control from secure environments (see Elliot, 2005) will not increase disclosure risk. This is because the original source data is not publicly available and therefore the research code cannot lead to the identification of individual or other infringements of privacy. This is consistent with the FAIR principles that meta-data associated with research data should be Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). All code and syntax should be reusable on later versions of a micro-level administrative social science data set either directly or with minor modification to enable the rerunning of the analysis.

At the current time research code is occasionally made available, but this practice is piecemeal. In the next section we illustrate some examples of code sharing and argue that current practices are suboptimal. Drawing on insights from computer science and other disciplines that have been engaged in e-Research we will suggest how some tools and environments would better support sharing code.

Current research code and syntax sharing practices in social science

Research code and syntax files are not routinely shared in the social sciences. There are existing examples of research code and syntax files being shared to undertake specific tasks, for example the creation of a social science measure.[23] There are also examples of research code being shared for complete projects.[24] One way that syntax files are made available is on personal websites. There are a number of leaders in specific social science fields whose websites provide key resources for other data analysts. Three examples from sociology include Professor Donald Treiman,[25] Professor Harry Ganzeboom[26] and Professor David Grusky and his collaborators.[27] A very good example of sharing research code for use with administrative micro-level data is the Wiki Space personally developed by Dr Rebecca Allen for the National Pupil Database.[28] This resource provides structured listings of code in relation to specific research areas (e.g. ethnicity) using a micro-level administrative social science data set.

Syntax files published on websites provide valuable assets, but we consider that this practice is suboptimal as a general mechanism for social science code sharing for a number of reasons. First, personal webpages are always at risk of going offline (e.g. the researcher moves institution). Second, there is no persistent record of the resources that personal web pages have contained, and there is seldom an audit trail of changes and updates. Third, the resources provided on a webpage are sometimes intrinsically linked to a specific project and are often not updated beyond the lifetime of the project. Fourth, updating and maintaining the resources provided on a website usually relies on the goodwill of the author. Fifth, the resources provided on webpages are often difficult to locate even with modern search engines, and are often only known by other researchers working in cognate social science areas. Sixth, data stored on personal webpages cannot easily be cited.[29]

There are examples of web-based facilities that have been specifically funded to support sharing research code relating to social science data analysis (see Lambert, 2015). Unfortunately, the longevity of these resources has been patchy and there are notable examples of facilities that are in decline or that have fallen into disrepair (for example they have broken links or have not recently been updated). Examples include two notable services that were funded by the

UK ESRC. The Grid Enabled Specialist Data Environment Services[30] which have not been functional for a couple of years, and Methodbox[31] whose 'current events' page has not been updated since 2012. Some progress has been made in making web-based resources more sustainable. The UK ESRC funded project ReStore[32] was specifically tasked with providing and curating a more sustainable web repository.

There are a number of repositories which are designed to share data and analyses (especially working papers, but also journal articles). Repositories are more common in economics and psychology compared with other social science disciplines.[33] DataCite[34] provide a registry of repositories through the re3data[35] initiative. Notable examples include the Harvard Dataverse Network,[36] Interuniversity Consortium for Political and Social Research,[37] Figshare,[38] Psych File Drawer,[39] REPEC[40] and DRYAD.[41] The Statistical Software Components[42] archive is a REPEC repository containing user-written software for statistical data analysis in a number of computer languages (but mostly in Stata). The resources produced by Professor John Hendrickx provide a notable example of research code sharing using REPEC.[43]

Research code published in the existing repositories provide valuable assets, and we consider that sharing code to support reproducible analysis of micro-level administrative social science data using existing repositories would be a step in the right direction. The existing repositories are likely to be long-lasting and this removes the problem of research being shared on webpages and then sites going offline. The existing repositories are easy to navigate to, and comparatively more 'searchable' than isolated webpages. Some repositories, such as the Harvard Dataverse, create a citation and Digital Object Identifier (DOI) for each set of replication materials, which provides an incentive to researchers to upload such materials. Whilst the Harvard Dataverse does permit users to record changes to uploaded information and code, this is rare among current repositories. In the next section we describe how insights from computer science may help improve upon existing practices.

Insights from computer science and e-research

Version control software (VCS) has been used extensively in professional software engineering to manage changes to files (Cochez et al., 2013). In practice, these technologies enable multiple people to work together collaboratively whilst also permitting flexible change tracking, and the ability to revert to previous versions (Cochez et al., 2013; Sink, 2011). Centralised version control software retains the development repository on a central server, a popular example being Apache Subversion.[44] Distributed version control software enables users to keep a local copy of the repository which can be synchronized with a master repository (Muşlu et al., 2014).

The use of VCS has been recommended in computer science (Peng, 2011; Stodden and Miguez, 2014), behavioural science (Adolph et al., 2012) and cognitive neuroscience (Yarkoni et al., 2010). Gentzkow and Shapiro (2014) is a rare example of VCS being recommended in the social sciences. VCS is particularly well suited to text files rather than binary or data files (Ram, 2013). Therefore it is appropriate for the syntax files that are used in micro-level administrative social science data analyses. Git and Mercurial are examples of popular distributed VCS environments. To support sharing, online code repository hosting services such as GitHub[45] and BitBucket[46] are widely used.

A Git repository is a database containing all the information needed to retain and manage the revisions and history of a project.[47] There are a number of recognisable advantages in using a protocol like Git to share research code for micro-level administrative social science data research. First, the version control philosophy in computing science chimes with the idea of the workflow in social science data analysis. Second, the Git environment provides an audit trail. Third, the ability to 'roll backwards' to previous versions of code provides an efficient facility when developing analysis. Fourth, if shared on an open online code repository, code sharing is automated and this enhances both transparency and reproducibility and it enables others to update or maintain existing syntax files and removes the burden from the original author. Fifth, micro-attribution (i.e. crediting researchers for their contributions) is automated through the use of inbuilt tools.

An impressive example of an attempt to make a complete project reproducible is Boring et al. (2016) which uses Git to make both data and methods reproducible.[48] Another innovative example of the possibilities for reproducible research is the publication of the methodology and code supporting the BuzzFeed News/BBC article, 'The Tennis Racket', which was published on 17 January 2016.[49] These two examples convince us of the utility of a Git approach for code sharing and its potential for micro-level administrative social science data research.

We are not arguing that the use of VCS is a panacea. There is always a time-cost associated with learning to use a new piece of software and to learning to work within a new computing environment. The time commitment will generally be greater for social science researchers who are less computationally able. VCS packages have been designed for use by software engineers and are not immediately welcoming to social science users. We are encouraged by initiatives such as

Software Carpentry[50] which provides workshops to help people without software engineering training to get the most out of the tools and techniques that are considered best practice for software development (see Wilson, 2006, 2014). It is reported that these activities have proven to be very successful in some scientific areas (Goble, 2014; Wilson et al., 2014). There are also numerous websites which have been created to assist researchers in understanding how to use distributed VCS. For example a proposal of good practice for working with Git is provided in Vincent Driessen's post.[51] An excellent introduction to the use of VCS in social sciences is posted by Andrew Hardie[52] and by Carly Strasser from the California Digital Library.[53]

Bird et al. (2009) describe a series of technical perils associated with using Git which should not be overlooked when considering its use in social science research. These perils are related to using Git in an unstructured manner and arise from disorganised working practices. On reflection these potential dangers can be mitigated if Git is used as a principled and organised aspect of the social science research workflow. Using Git is not a remedy for an *ad-hoc*, poorly planned and inadequately documented workflow. A planned workflow is integral to undertaking reproducible research using micro-level administrative social science data.

There are a number of emerging initiatives which seek to package aspects of the research process to record and to make all the files associated with a project visible. These approaches integrate VCS with more aspects of the research project, including meta-data, data, code (syntax files), and documentation.[54] Research Objects[55] (ROs) are a means to package up research outputs (data, metadata, code, results, documentation, papers, etc.) for describing and associating resources.

> 'An RO bundles together essential information relating to experiments and investigations. This includes not only the data used, and methods employed to produce and analyse that data, but also the people involved in the investigation.' (Bechhofer et al., 2013: 600)

This is achieved using a tool to create a RO and associate files with it, for example RO Manager.[56] This enables users of ROs to gain access to reproducible research work (Bechhofer et al., 2013; Belhajjame et al., 2012; Hettne et al., 2014). VCS is central to the building of ROs. Another system which allows the upload of data and code, which can then be shared (and cited) is the Open Science Framework. This framework incorporates version control and represents an alternative but broadly comparable system to ROs. These systems are entirely consistent with the FAIR principles for data and metadata described earlier in this paper and would be a mechanism for associating the code used to extract and link data sets with the syntax files used to prepare and analyse the micro-level administrative social science data sets.

We have recently become aware of an interesting initiative in this area, the Farr Commons project. We understand that it will develop a ROs framework which aims to create an infrastructure to enable members of the UK Farr Institute (which is a health informatics collaboration closely related to UK ADRN) to easily and securely, share and reuse methodology and data (see Pavis and Morris, 2015).[57] Our understanding is that it is a pilot for the NIH RO Commons, with a stated aim of describing a set of rules for contributing to the data commons, which enable correct identification of ROs.[58] The commons are described as a conceptual framework for a digital environment to allow efficient storage, manipulation and sharing of ROs.[59] The technology and practices associated with creating ROs are currently emerging. It is clear to us that the potential advantages of such technologies will only be reaped if the administrative social science data analysis community begin to organise their research endeavours in systematic and organised fashions, for example which are supported by VCS.

Insights from other disciplines

There are further methods that other disciplines have employed to improve reproducibility. Focusing largely on the fields of medicine and psychology, this section briefly summarises these practices.

Substantial efforts have been made to improve transparency and reproducibility in clinical trials. Mathieu et al. (2009) compared pre-registered study protocols with published study outcomes and identified that selective reporting and lack of adequate registration are prevalent. Chan et al. (2014) argues that dissemination of research protocols, reports and individual-level data sets is instrumental in improving reproducibility. This also needs to be matched by the adoption of consistent standards for protocols and rewards for compliance with these practices by academic institutions, journals and research funders (Chan et al., 2014). The applications process to access administrative social science data sets may help improve reproducibility. This is because the research questions, data sources and methods to be employed must be specified prior to data access being granted.

In the field of psychology, work has been undertaken to estimate reproducibility through systematic reviews. Open Science Collaboration (2015) replicated 100 experimental and correlational studies and concluded that the strength of the original evidence in the studies they reviewed was the greatest prediction of successful replication. They also argued that the improvements to

the quality and credibility of the literature supporting scientific work through initiatives (such as TOP) were important steps for achieving reproducibility.

A crucial aspect for replication is suitable meta-data accompanying the data resources available. Data resource profiles (such as those published by the International Journal for Epidemiology) are examples of good practice. For instance, in the field of epidemiology, administrative data resource profiles have been published for the Children Looked After Return in England (Mc Grath-Lone et al., 2016), the Scottish National Prescribing Information System (Alvarez-Madrazo et al., 2016), the Scottish Longitudinal Study (Boyle et al., 2009), and the Swedish Microdata Research from Childhood into Lifelong Health and Welfare data sets (Lindgren et al., 2016). These profiles are invaluable sources of information with respect to replication, particularly in understanding the characteristics and features of the data sets being analysed.

Repositories recording studies using primary care data have been collated by organisations such as Clinical Practice Research Datalink[60] and The Health Improvement Network.[61] These collections are a vital component in reproducible research using health data as researchers can review other work using similar data and learn more about the data resources. The REporting of studies Conducted using Observational Routinely-collected Data[62] issues a checklist for the items that should be reported in observational studies using routinely collected health data. Initiatives such as the Enhancing the QUAlity and Transparency Of health Research (EQUATOR)[63] offer a library of reporting guidelines to further aid consistency. This helps ensure transparency and consistency and ultimately replicability (Benchimol et al., 2015).

Conclusion

The expansion of administrative data that were originally collected to organise, manage, monitor or deliver services but which are suitable for social science, offers exciting research prospects. The previous restrictions on access to administrative data are beginning to be lifted in a number of nations. In the UK, the ADRN has been created to improve access to micro-level administrative social science data. Social scientists will gain carefully supervised access to previously unavailable data from government departments and other agencies.

Administrative social science data have not primarily been collected for social science research. In practice administrative social science data sets will often only include a restricted set of social science related explanatory variables (compared with a large-scale social survey which has the primary goal of collecting social science data). Administrative data are collected in order to organise, manage, monitor or deliver services and may not be organised in the most optimal structure for social science research. Measures collected in administrative data sets are usually of variable quality. For many research questions data will be required from a number of different sources which have to be linked together. The accuracy of the process of linking records can vary. In many circumstances better developed meta-data would make a positive contribution. Considered together these issues indicate the practical messiness of administrative data, and illustrate the complexity of undertaking social science analyses using micro-level administrative data.

It is difficult to contrive an argument for research not being reproducible. Research data should never be destroyed because this makes it impossible to reproduce research findings. It is similarly difficult to imagine situations in which accuracy, efficiency and transparency were not desirable features of research. Not having a planned and organised workflow and not using syntax when analysing micro-level administrative social science data can be compared to drinking and driving. In both cases it doesn't matter how careful you are, it is still highly likely to end in a wreck![64] Therefore just like drinking and driving, we strongly warn against this practice. No researcher should ever analyse micro-level administrative social science data without a planned and organised workflow that uses syntax. The 'take home' message is that reproducibility should be taken seriously.

We anticipate that the UK ADRN will have a leading role in helping the data production community to navigate towards best practices. We envisage that in the near future companion work on reproducibility in micro-level administrative social science data construction will be published that sets guidelines that are feasible given the current practical, technological, political, legal and ethical issues that surround the production of research data sets.

Most readers will have a fond, or possibly even a terrifying, early educational memory of being told to 'show their working out'. Somewhere between elementary school and graduate school this requirement has become more relaxed. In a nutshell we believe that enough information to check that results are correct and that conclusions are plausible should be provided. This should be accompanied by enough information to describe which analyses were intended, and which were actually undertaken. This transparency should be achieved through sharing research code, by which we mean the publication of adequately annotated syntax files and other supporting research documentation. We therefore advocate that researchers are allowed to export syntax that has been suitably cleared through

the normal protocol of statistical disclosure control from secure environments (see Elliot, 2005). This will enable its reuse and scrutiny by the research community.

Diggle (2015: 808) states that:

> 'In many scientific areas, most obviously the health sciences, concern about preserving the confidentiality of information on human subjects needs to be balanced against the public benefit of insightful statistical analysis (and sometimes critical reanalysis) of disaggregated data.'

We argue that such 'critical reanalysis' should be considered as being part of the benefit to the public. This point is reinforced in public consultations on the use of administrative data, where some respondents recognised the importance of data retention for a period sufficient to ensure that analysis could be reproduced (Cameron et al., 2014: 47). Where it is legal we would advocate sharing research data along with research code.[65]

At the current time we suggest that much progress can be made by adopting the version control philosophy in computing science and e-Research, which accords with the idea of the workflow in social science data analysis. Git repositories provide a useable and effective solution to 'research code sharing' in administrative social science data research. In particular, the Git environment provides an audit trail, but also supports the ability to 'roll backwards' to previous versions of code. Sharing can be automated within the Git environment through global repositories such as GitHub and BitBucket and this enhances transparency, maintainability, and ultimately reproducibility.

Collaboration is central to research code sharing and most likely to encourage code sharing when contributors are appropriately acknowledged. The Git environment is well suited to micro-attribution through its automated use of inbuilt tools. Finally, the Git environment provides an essential stepping stone to emerging technologies such as ROs that provide packaged up research results, containing syntax, data (when permitted), metadata and documentation, for others to reproduce analyses and build upon research.

Acknowledgements

We gratefully acknowledge the comments of Professor Philip Stark (UC Berkeley), Kellie Ottoboni (UC Berkeley), Dr Nicole Janz (University of Cambridge), Dr Alasdair Rutherford (University of Stirling), Dr Iain Atherton (Edinburgh Napier University), and Dr Kevin Ralston (University of Edinburgh). We would also like to thank Professor Chris Dibben (University of Edinburgh) and colleagues at the Administrative Data Research Centre – Scotland.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Economic and Social Research Council for Administrative Data Research Centre – Scotland project under grant number [ES/L007487/1].

Notes

1. See https://cos.io/top/ (accessed 25 February 2016).
2. See https://osf.io/2cz65/?_ga=1.69210640.1492415597.1457357091 (accessed 7 March 2016).
3. See http://www.bitss.org/ (accessed 13 May 2016).
4. See http://adrn.ac.uk/ (accessed 24 February 2016).
5. There are many non-personal administrative data sets available but the ADRN has been created to assist with access to individual-level data. The UK government produces a list of administrative data sources for each government department, see http://www.adls.ac.uk/find-administrative-data/official-statements-of-administrative-sources/ (accessed 15 September 2016).
6. For an overview of the ADRN, please see https://www.youtube.com/watch?v=E3e4D2bHxa8 (accessed 26 April 2016).
7. An example of aggregate-level administrative data is the number of births in Scotland by council area (see http://statistics.gov.scot/data/births accessed 15 September 2016). The distinction between macro- and micro-level administrative data is required because these data have different characteristics, different access procedures and different uses within social science research.
8. The variable by case matrix will be familiar to researchers who have been trained to undertake statistical analyses of social science data, and it is described in standard elementary textbooks, for example see De Vaus (2014).
9. We use the term e-Research as a label for large-scale science that is increasingly being carried out through distributed global collaborations which typically feature scientific enterprises that require access to very large data collections and large-scale and high performance computing resources.
10. We recognise that researchers use different statistical software packages but that the principle of using syntax files is generic.
11. For a list of UK facilities, see https://adrn.ac.uk/protecting-privacy/secure-environment/safe-centres/ (accessed 26 April 2016).
12. See http://www.statcan.gc.ca/pub/75f0011x/2013001/serv-eng.htm (accessed 24 February 2016).
13. See https://osf.io/9f6gx/wiki/Guidelines/ (accessed 7 March 2016).
14. See https://adrn.ac.uk/media/1169/adrn-034-data-retention_pub.pdf (accessed 19 September 2016).

15. See http://www.diw.de/en/soep (accessed 24 February 2016).
16. See https://www.iser.essex.ac.uk/bhps/documentation (accessed 24 February 2016).
17. See http://www.nesstar.com/about/about.html (accessed 24 February 2016).
18. See http://sls.lscs.ac.uk/variables/ (accessed 24 February 2016).
19. See https://www.gov.uk/government/publications/national-pupil-database-user-guide-and-supporting-information (accessed 24 February 2016).
20. Government open data resources have been published online, see https://data.gov.uk/about (accessed 26 April 2016). However, these do not included micro-level social science administrative data sets because of legal constraints.
21. For example the UK Data Archive (http://www.data-archive.ac.uk/, accessed 11 May 2016) or the Inter-university Consortium for Political and Social Research Data Repository (http://www.icpsr.umich.edu/index.html, accessed 11 May 2016).
22. See https://adrn.ac.uk/protecting-privacy/de-identified-data/trusted-third-parties (accessed 2 May 2016).
23. See http://www.camsis.stir.ac.uk (accessed 24 February 2016).
24. See http://www.restore.ac.uk/Longitudinal/qv/ (accessed 24 February 2016).
25. See http://www.ccpr.ucla.edu/dtreiman (accessed 24 February 2016).
26. See http://www.harryganzeboom.nl/index.htm (accessed 24 February 2016).
27. See http://www.classmobility.org/ (accessed 24 February 2016).
28. See https://nationalpupildatabase.wikispaces.com/ (accessed 24 February 2016).
29. There is an initiative to improve citation of source code, see https://www.force11.org/software-citation-principles (accessed 12 May 2015).
30. See http://www.dames.org.uk/ (accessed 24 February 2016).
31. See www.methodbox.org (accessed 24 February 2016).
32. See http://www.restore.ac.uk/ (accessed 24 February 2016).
33. A useful short summary and links are provided by the Berkeley Initiative for Transparency in the Social Sciences, see http://www.bitss.org/resource-tag/data-repository/ (accessed 7 March 2016).
34. See https://www.datacite.org/ (accessed 24 February 2016).
35. See http://service.re3data.org/about (accessed 26 April 2016).
36. See https://thedata.harvard.edu/dvn/ (accessed 24 February 2016).
37. See https://www.icpsr.umich.edu/icpsrweb/landing.jsp (accessed 24 February 2016).
38. See http://figshare.com/ (accessed 24 February 2016).
39. See http://psychfiledrawer.org/ (accessed 24 February 2016).
40. See http://repec.org/ (accessed 24 February 2016).
41. See http://datadryad.org/ (accessed 7 March 2016).
42. See http://fmwww.bc.edu/repec/bocode/s/sscstats.html (accessed 24 February 2016).
43. See https://ideas.repec.org/e/phe38.html (accessed 24 February 2016).
44. Apache Subversion is often abbreviated SVN, after the command *svn*.
45. See https://github.com/ (accessed 22 February 2016).
46. See https://bitbucket.org/ (accessed 22 February 2016).
47. Loeliger and McCullough (2012) provide an excellent description of the terminology used when working with Git.
48. See https://github.com/kellieotto/SET-and-Gender-Bias (accessed 24 February 2016).
49. See http://www.buzzfeed.com/heidiblake/the-tennis-racket (accessed 24 February 2016). The publically available data and the code used in the analyses are shared via a GitHub repository, see https://github.com/BuzzFeedNews/2016-01-tennis-betting-analysis (accessed 24 February 2016).
50. See http://software-carpentry.org/ (accessed 24 February 2016).
51. See http://nvie.com/posts/a-successful-git-branching-model/ (accessed 24 February 2016).
52. See http://cass.lancs.ac.uk/?tag=version-control-software (accessed 24 February 2016).
53. See http://datapub.cdlib.org/2014/05/05/github-a-primer-for-researchers/ (accessed 24 February 2016).
54. For a technical and conceptual computer science introduction to how ROs might be implemented, see http://www.slideshare.net/matthewgamble/introduction-to-research-objects-cw2015 (accessed 24 February 2016).
55. See http://www.researchobject.org/overview/ (accessed 24 February 2016).
56. See https://github.com/wf4ever/ro-manager (accessed 24 February 2016).
57. See http://farrcommons.github.io/ (accessed 24 February 2016).
58. See http://farrcommons.github.io/rules.html (accessed 24 February 2016).
59. See http://farrcommons.github.io/ (accessed 24 February 2016).
60. See https://www.cprd.com/intro.asp (accessed 19 September 2016).
61. See https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database (accessed 19 September 2016).
62. See http://www.record-statement.org/ (accessed 19 September 2016).
63. See http://www.equator-network.org/ (accessed 19 September 2016).
64. We are grateful to Professor Philip Stark, University of California Berkeley, for this useful and clear analogy.
65. For a discussion of sharing synthetic administrative data, please see http://www.vernongayle.com/blog-research.html (accessed 24 October 2016).

References

Adolph KE, Gilmore RO, Freeman C, et al. (2012) Toward open behavioral science. *Psychological Inquiry* 23: 244–247.

Alvarez-Madrazo S, McTaggart S, Nangle C, et al. (2016) Data resource profile: The Scottish National Prescribing Information System (PIS). *International Journal of Epidemiology* 45(3): 714–715f.

Bechhofer S, Buchan I, De Roure D, et al. (2013) Why linked data is not enough for scientists. *Future Generation Computer Systems* 29: 599–611.

Belhajjame K, Corcho O, Garijo D, et al. (2012) Workflow-centric research objects: First class citizens in scholarly discourse. In: *SePublica2012 at ESWC2012*, 28 May 2012, Greece.

Benchimol EI, Smeeth L, Guttmann A, et al. (2015) The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 12: e1001885.

Bird C, Rigby PC, Barr ET, et al. (2009) The promises and perils of mining git. In: *Mining software repositories, 2009. MSR '09. 6th IEEE international working conference*, The Westin Bayshore, Vancouver, BC, Canada, 16–17 May 2009, pp. 1–10.

Blum JM, Warner G, Jones S, et al. (2009) Metadata creation, transformation and discovery for social science data management: The DAMES Project infrastructure. *IASSIST Quarterly* 33: 23–30.

Boring A, Ottoboni K and Stark PB (2016) Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen*. DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Boslaugh S (2005) *An Intermediate Guide to SPSS Programming: Using Syntax for Data Management*. Thousand Oaks, CA: Sage Publications.

Boyle PJ, Feijten P, Feng Z, et al. (2009) Cohort profile: The Scottish Longitudinal Study (SLS). *International Journal of Epidemiology* 38: 385–392.

Cameron D, Pope S and Clemence M (2014) *Dialogue on Data: Exploring the Public's Views on Using Administrative Data for Research Purposes*. London: Ipsos Mori Social Research Institute.

Card D, Chetty R, Feldstein MS, et al. (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*. Available at: http://ssrn.com/abstract=1888586 or http://dx.doi.org/10.2139/ssrn.1888586 (accessed 24 October 2016).

Chan A-W, Song F, Vickers A, et al. (2014) Increasing value and reducing waste: Addressing inaccessible research. *The Lancet* 383: 257–266.

Cochez M, Isomöttönen V, Tirronen V, et al. (2013) How do computer science students use distributed version control systems? In: Ermolayev V, Mayr H, Nikitchenko M, et al. (eds) *Information and Communication Technologies in Education, Research, and Industrial Applications* Cham (ZG), Switzerland: Springer International Publishing, pp. 210–228.

Connelly R, Playford CJ, Gayle V, et al. (2016) The role of administrative data in the Big Data revolution in social science research. *Social Science Research* 59: 1–12.

De Vaus DA (2014) *Surveys in Social Research*. Abingdon, Oxon: Routledge.

Dibben C, Elliot M, Gowans H, et al. (2015) The data linkage environment. *Methodological Developments in Data Linkage*. Chichester: John Wiley & Sons, Ltd, pp. 36–62.

Diggle PJ (2015) Statistics: A data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178: 793–813.

Doiron D, Raina P and Fortier I (2013) Linking Canadian population health data: Maximizing the potential of cohort and administrative data. *Canadian Journal of Public Health* 104: e258–e261.

Elias P (2014) Administrative data. In: Duşa A, Nelle D, Stock G and Wagner GG (eds) *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. Berlin: Scivero Verlag, p. 47.

Elliot MJ (2005) Statistical disclosure control. *Encyclopedia of Social Measurement*. New York, NY: Elsevier, pp. 663–670.

Figlio DN, Karbownik K and Salvanes KG (2015) Education research and administrative data. *National Bureau of Economic Research Working Paper Series No. 21592*.

Freese J (2007) Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research* 36: 153–172.

Gentzkow M and Shapiro JM (2014) *Code and data for the social sciences: A practitioner's guide,* University of Chicago mimeo. Available at: https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf (accessed 13 December 2016).

Giles P (2001) An overview of the Survey of Laobur and Income Dynamics (SLID). *Canadian Studies in Population* 28: 363–375.

Goble C (2014) Better software, better research. *Internet Computing, IEEE* 18: 4–8.

Hanson B, Sugden A and Alberts B (2011) Making data maximally available. *Science* 331: 649.

Harford T (2014) Big Data: A big mistake? *Significance* 11: 14–19.

Hettne K, Dharuri H, Zhao J, et al. (2014) Structuring research methods and data with the research object model: Genomics workflows as a case study. *Journal of Biomedical Semantics* 5: 41.

Hey AJG, Tansley S and Tolle KM (2009) *The Fourth Paradigm Data-intensive Scientific Discovery (Version 1.1)*. Redmond, WA: Microsoft Research.

Holman CD, Bass AJ, Rosman DL, et al. (2008) A decade of data linkage in Western Australia: Strategic design, applications and benefits of the WA data linkage system. *Australian Health Review* 32: 766–777.

Jahnke L, Asher A and Keralis SDC (2012) *The Problem of Data*. Washington, DC: Council on Library and Information Resources.

Janz N (2015) Bringing the gold standard into the classroom: Replication in University Teaching. *International Studies Perspectives*. DOI: 10.1111/insp.12104.

King G (1995) Replication, replication. *PS: Political Science and Politics* 28: 443–499.

King G (2003) The future of replication. *International Studies Perspectives* 4: 100–105.

King G (2011) Ensuring the data-rich future of the social sciences. *Science* 331: 719–721.

Lambert PS (2015) Advances in data management for social survey research. In: Halfpenny P and Procter R (eds) *Innovations in Digital Research Methods*. London: Sage, pp. 105–122.

Lambert PS and Gayle V (2008) Data management and standardisation: A methodological comment on using results from the UK Research Assessment Exercise 2008. *Technical Paper* 3.

Lindgren U, Nilsson K, de Luna X, et al. (2016) Data resource profile: Swedish Microdata Research from Childhood into Lifelong Health and Welfare (Umeå SIMSAM Lab). *International Journal of Epidemiology* 45(4): 1075–1075g.

Loeliger J and McCullough M (2012) *Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development*. Sebastopol, CA: O'Reilly Media, Inc.

Long JS (2009) *The Workflow of Data Analysis using Stata*. College Station, TX: Stata Press.

Long JS and Freese J (2014) *Regression Models for Categorical Dependent Variables Using Stata*. College Station, TX: Stata Press.

Marsh C and Elliott J (2008) *Exploring Data: An Introduction to Data Analysis for Social Scientists*. Cambridge: Polity Press.

Mathieu S, Boutron I, Moher D, et al. (2009) Comparison of registered and published primary outcomes in randomized controlled trials. *Journal of American Medical Association* 302: 977–984.

Mc Grath-Lone L, Harron K, Dearden L, et al. (2016) Data resource profile: Children looked after return (CLA). *International Journal of Epidemiology* 45: 716–717.

McCullough BD, McGeary KA and Harrison TD (2008) Do economics journal archives promote replicable research? *The Canadian Journal of Economics/Revue canadienne d'Economique* 41: 1406–1420.

Muşlu K, Bird C, Nagappan N, et al. (2014) Transition from centralized to decentralized version control systems: A case study on reasons, barriers, and outcomes. In: *Proceedings of the 36th international conference on software engineering,* ACM, 4 June 2014, pp. 334–344.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 349(6251): aac47161–8.

Organisation for Economic Co-operation and Development (2007) *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD.

Pavis S and Morris AD (2015) Unleashing the power of administrative health data: The Scottish model. *Public Health Research & Practice* 25: e2541541.

Peng RD (2011) Reproducible research in computational science. *Science* 334: 1226–1227.

Ram K (2013) Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine* 8: 7.

'Reality check on reproducibility' [Editorial] (2016) *Nature* 533: 437.

Sandve GK, Nekrutenko A, Taylor J, et al. (2013) Ten simple rules for reproducible computational research. *PLoS Computational Biology* 9: e1003285.

Sink E (2011) *Version Control by Example*. Champaign, IL: Pyrenean Gold Press.

Stodden V and Miguez S (2014) Best practices for computational science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software* 2(1): e21, 1–6.

Treiman DJ (2009) *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco, CA: Jossey-Bass.

Tukey JW (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wallgren A and Wallgren B (2007) *Register-based Statistics: Administrative Data for Statistical Purposes*. Chichester: John Wiley & Sons.

Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

Wilson G (2006) Software carpentry: Getting scientists to write better code by making them more productive. *Computing in Science & Engineering* 8: 66–69.

Wilson G (2014) Software carpentry: lessons learned [version 1; referees: 3 approved]. *F1000Research* 3: 62.

Wilson G, Aruliah DA, Brown CT, et al. (2014) Best practices for scientific computing. *PLoS Biology* 12: e1001745.

Woollard M (2014) Administrative data: Problems and benefits. A perspective from the United Kingdom. In: Duşa A, Nelle D, Stock G, et al. (eds) *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. Berlin: SCIVERO Verlag, p. 49.

Yale Law School Roundtable on Data and Code Sharing. Reproducible research. *Computing in Science & Engineering* 12: 8–13.

Yarkoni T, Poldrack RA, Van Essen DC, et al. (2010) Cognitive neuroscience 2.0: Building a cumulative science of human brain function. *Trends in Cognitive Sciences* 14: 489–496.