OPEN LETTER

REVISED Navigating the landscape of non-health administrative data in Scotland: A researcher's narrative [version 2; peer review: 2 approved]

Previously titled: Navigating the landscape of administrative data in Scotland

Matthew H. Iveson [1-3], Ian J. Deary [1]

[1]Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK
[2]Administrative Data Research Centre Scotland, Edinburgh, UK
[3]Centre for Clinical Brain Sciences, The University of Edinburgh, Edinburgh, UK

v2 **First published:** 17 Jun 2019, **4**:97 (https://doi.org/10.12688/wellcomeopenres.15336.1)
Latest published: 21 Oct 2019, **4**:97 (https://doi.org/10.12688/wellcomeopenres.15336.2)

Abstract

Background: There is growing interest in using routinely collected administrative data for research purposes. Following the success of research using routinely collected healthcare data, attention has turned to leveraging routinely-collected non-health data derived from systems providing other services to the population (e.g., education, social security) to conduct research on important social problems. In Scotland, specialised organisations have been set up to support researchers in their pursuit of using and linking administrative data. The landscape of administrative data in Scotland, however, is complex and changeable, and is often difficult for researchers to navigate.
Purpose: This paper provides a researcher's narrative of the steps required to gain the various approvals necessary to access and link non-health administrative data for research in social and cognitive epidemiology.
Findings: This paper highlights the problems, particularly regarding the length and complexity of the process, which researchers typically face, and which result in a challenging research environment. The causes of these problems are discussed, as are potential solutions.
Conclusions: Whereas the potential of non-health administrative data is great, more work and investment are needed on the part of all those concerned – from researchers to data controllers – in order to realise this potential.

Keywords

Administrative data, Big data, Data linkage, Narrative, Social epidemiology

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
REVISED **version 2** published 21 Oct 2019	✓ report	✓ report
version 1 published 17 Jun 2019	✓ report	? report

1 **Chris Playford**, University of Exeter, Exeter, UK

2 **Michael Fleming**, University of Glasgow, Glasgow, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Matthew H. Iveson (Matthew.Iveson@ed.ac.uk)

Author roles: Iveson MH: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Deary IJ**: Conceptualization, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by a UK cross council Lifelong Health and Wellbeing Initiative (grant number: MRCG1001401), for which I.J.D. is the principal investigator. M.H.I. and I.J.D. are members of The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (grant number: MR/K026992/1). I.J.D. is part of the "Stratifying Resilience and Depression Longitudinally" (STRADL) project, which is funded by the Wellcome Trust through a Strategic Award (grant number: 104036). M.H.I. is a member of the Administrative Data Research Centre Scotland, supported by the Economic and Social Research Council (grant number: ES/L007487/1), and is part of an Medical Research Council Mental Health Data Pathfinder project (grant number: MRC - MC_PC_17209).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Iveson MH and Deary IJ. **Navigating the landscape of non-health administrative data in Scotland: A researcher's narrative [version 2; peer review: 2 approved]** Wellcome Open Research 2019, **4**:97 (https://doi.org/10.12688/wellcomeopenres.15336.2)

First published: 17 Jun 2019, **4**:97 (https://doi.org/10.12688/wellcomeopenres.15336.1)

REVISED **Amendments from Version 1**

The updated text includes changes suggested by the reviewers, as well as small updates based on developments to the administrative data landscape. We have changed the title of the manuscript and several terms within to better reflect the focus on non-health administrative data, though we have also better discussed the development of health administrative data as a comparison. Several references have been added as helpfully suggested by the reviewers, and we have used these to develop our discussion of the issues surrounding administrative data access. Finally, we have produced a more balanced picture of the data access process by discussing the purpose of procedures such as information governance, clarifying that any attempt to improve non-health administrative data access should not come at the cost of data privacy.

Any further responses from the reviewers can be found at the end of the article

Introduction

The rise of big data represents a revolutionary opportunity for both researchers and policy makers. This opportunity has been perhaps best recognised by Scandinavian countries (Sweden, Norway and Denmark), in which national databases – including healthcare and conscription data – have been linked together using unique personal identification numbers, allowing for large and powerful research studies[1]. These studies have significantly improved our understanding of issues such as cancer[2], mental health conditions[3], pre-term birth[4], cognitive ageing[5], socioeconomic inequality[6], etc. In Scotland, previous work has already leveraged routinely collected health-related administrative data, such as that from the National Health Service, to address questions regarding how morbidity and mortality relate to people's social background and psychological differences[7–10]. Health data research has benefitted from increasing investment (from both governments and research councils), and from several high-profile public promotions (e.g., the 'data saves lives' campaign). In the last decade, researchers have extended their sights to routinely collected administrative data, such as that from the Scottish Government, as a largely untapped resource with similar potential for impact and societal benefit. These requests have been facilitated by purpose-built organisations such as the Administrative Data Research Centre Scotland (funded by the ESRC). The role of these new organisations is to support researchers and to negotiate access to both health and non-health administrative data on their behalf. Furthermore, many of the organisations controlling non-health administrative data have begun to develop and implement processes for dealing with data requests. In contrast to earlier efforts, then, data access and linkage for research purposes should be easier and faster. However, despite the promise of non-health administrative data, the road to obtaining data is not always smooth.

Below we give a researcher's perspective on the journey through the landscape of administrative data in Scotland. The narrative describes and comments on a project devised to link the Scottish Mental Survey 1947 cohort (SMS1947)[11] to routinely collected health and non-health administrative data, including the Scottish Census. This follows on from and extends similar efforts to link the same cohort to routinely-collected health administrative data, carried out before major changes in the Scottish landscape of big data[8–10,12]. While previous efforts have used linked SMS1947 and health data to investigate life-course determinants of cause-specific mortality[9], the current project sought to extend this linkage to non-health administrative datasets and use them to examine health and social care outcomes. The present account, then, is partly an update, now that data linkage organisations and processes in general are more mature, and also a major extension, given that access to non-health administrative data is a relatively recent development. The post-doctoral researcher employed as part of this project was in post for 26 months, from 1st August 2016 to 1st October 2018. We describe the process involved in acquiring and linking data for four specific studies (see Figure 1) – two involving data from the Scottish Census and two involving data from the

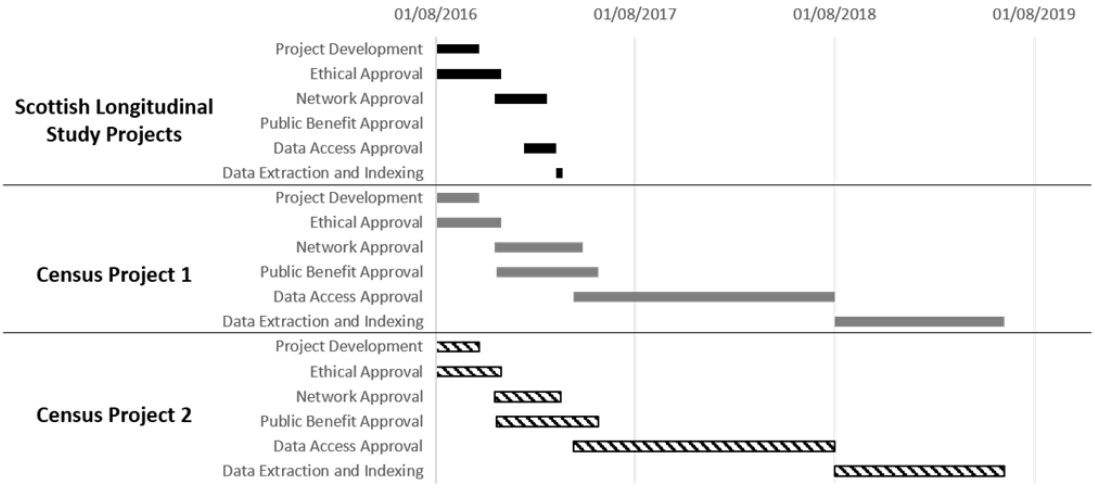

Figure 1. Gantt chart demonstrating the time taken to complete key stages for each project.

Scottish Longitudinal Study (SLS). The SLS is a standing resource containing pre-linked records for a 5.3% sample of the Scottish population, covering primarily census data, education data and hospitalisation data (see Box 1). As the SLS offers pre-linked data, data access approval and data extraction and indexing are much faster than non-SLS projects (Figure 1). The two SLS studies were conceived as interim studies to be conducted while completing the approvals process for the two census studies, and are included for comparison. The two census projects differed slightly in the specific datasets to be linked, though the cohort to be linked was the same. The organisations involved, their role, and they type of data they hold are summarised in Table 1. Note that we do not claim to be experts in the legalities and technicalities that motivate the information governance and data access processes described below; that is an unrealistic expectation to place on researcher. Instead, we describe the process as it is faced by researchers and reflect on the challenges that arise along the way. We provide this narrative in the hopes that it will inform researchers who are considering working with non-health administrative data, and that it will help to critically assess and improve current data governance and access policies.

Stage 1: Ethical approval (prep time: 3.5 months; processing time: 0.5 months)

The conception of the project began on 1st August 2016 (see Figure 1). Because pseudo-anonymised health data linkage had been achieved for the SMS1947 cohort as part of a

Box 1. Challenges to using the safe settings

Given the time required to arrange the linkage of the full Scottish Mental Survey 1947 cohort, we pursued projects using a sample of the cohort which had already been linked within the Scottish Longitudinal Study (SLS). As the SLS is a standing database of linked health and non-health administrative data, covering 5.3% of the Scottish population selected from 20 dates of birth, we reasoned that these projects would be quicker to obtain and analyse data. Note that the approval and access process is very different for SLS data than for data which has not been previously linked; procedures have been pre-agreed between data controllers and there is only a single point of contact. The SLS also provide research support functions to all users. Researchers using SLS data are required to do so in a 'safe setting' – a secure, monitored environment within Ladywell House (Edinburgh, UK). Data can only be viewed from this safe setting, and all analyses must be done on-site and later checked for potential statistical disclosure before being removed from the safe setting. We first visited the safe setting on 21st March 2017. Over the course of using the safe setting (442 days; from 21st March 2017 to 6th June 2018) we ran into several issues which lengthened the planned projects. Below we summarise these issues. Whereas some are specific to those projects using the Scottish Mental Survey 1947 cohort, the majority of the issues listed below reflect the type of trials faced by all users of the SLS safe setting, and of access-controlled data sites more generally (e.g., the ADRC-S safe havens).

Approval:

- Approval required for the project itself (1 form) and for the researchers themselves (2 forms)
- Additional aim required formal approval from SLS panel (submitted 11/04/2017). Approval was not communicated to the researchers until later (04/05/2017)

Availability:

- 3 delays in attendance due to SLS staff shortage/training

Changes in policy:

- Booking space in the safe setting changed to require 2 days' notice (12/10/2017)
- Intermediate output minimum cell count changed from 5 to 10 (02/11/2017)
- Disclosure control timelines changed from 5 working days for intermediate output and 15 working days for final output to 10 working days and 20 working days (29/03/2018)

Analyses:

- Initial dataset was missing a requested variable (21/03/2017). A new dataset was extracted (23/03/2017), including the missing variable, but was not made available until later (02/05/2017) due to staff shortage. A second missing variable was identified (25/05/2017) and was later added (12/06/2017).
- Analyses was conducted in R Studio using specialist packages[13]. These packages were not pre-installed on the safe setting machines, and needed to be requested (7 forms).
- Although the packages were installed, it emerged that their dependencies were not. These dependencies had to be subsequently requested (2 forms).
- The installed version of R Studio was not compatible with some of the installed packages, and so a newer version had to be requested (1 form)
- Some analyses were not included in intermediate statistical disclosure controlled output (18/05/2017)
- 4 intermediate outputs were redacted due to concerns over statistical disclosure. Concerns were raised particularly regarding the inclusion of Ns, despite these adhering to the disclosure control guidelines (all cells greater than 10, or censored accordingly).
- Concerns were also raised due to the cohort used – multiple projects working on the same cohort can produce tables which, when combined, are said to risk residual disclosure. Giving current researchers particular outputs may mean that future researchers are prevented from producing other outputs.

Table 1. Organisations, their role, and number of forms required. Merged cells indicate shared involvement.

Organisation	Role (Type of data held)	Forms submitted	Amendments submitted
Administrative Data Research Centre - Scotland	Research Support	0	0
ACCORD	Clinical Sponsor	1	0
NHS Research Ethics Committee	Ethical Approval		0
University of Edinburgh Legal Services	Institutional Guarantor	1	0
Administrative Data Research Network	Network Approval and Resources	5	0
Public Benefit and Privacy Panel	Public Benefit and Privacy Approval	2	5
NHS Information Services Division	Data Controller (Health data)	0	0
Electronic Data Research and Innovation Service	Research Coordinator for NHS ISD	0	0
National Records of Scotland	Data Controller (Births and Deaths data)/Trusted Third Party	4	0
Scottish Government	Data Controller (Census data)		0
Scottish Longitudinal Study	Data Resource (Pre-linked administrative data)	2	1

previous project[8–10], the first step was to determine whether the ethics and permissions obtained previously could be extended to cover the proposed project. Importantly, previous projects obtained specialised ethical approval due to the unconsented use and linkage of health administrative data. After several weeks of meetings and emails with colleagues in the Administrative Data Research Centre – Scotland (ADRC-S), NHS Information Services Division (NHS ISD), ACCORD (the clinical sponsor), and the NHS Research Ethics Committee (NHS REC), it was determined that new specialised approvals would need to be sought. This period reflected the relative unfamiliarity of some of these organisations with data linkage projects, and the conflicting interpretations of data linkage procedures. The researchers submitted an ethics application covering the new project to the NHS REC for initial review on 21st October 2016, and for final review on 18th November 2016. The ethics application was formally approved 11 days later, on 29th November 2016.

Stage 2: Network approval (prep time: 3 months; processing time: 0.5 months)

The next stage of the process was to obtain approval from the Administrative Data Research Network (ADRN) in order to be able to access their support and infrastructure. This involved the preparation of a second set of forms – one for each of the two census studies and one for the SLS studies within the project. We began preparing these forms on 17th November 2016 and submitted iterative versions to the ADRC-S for preliminary feedback on the 20th December 2016 and on 12th January 2017. Final versions of the forms were submitted to the ADRN in turn, from March through April of 2017. Approval for each study was obtained roughly 2 weeks after submission, with the final study being approved on 27th April 2017. Whereas these forms were to gain ADRN approval for the proposed studies, a third set of forms was required to gain ADRN approval for the researchers involved. This approval requires researchers to detail their research experience, to detail any previous incidences of data misuse, and to agree to abide by the ADRN's terms of use. Note that these forms were in addition to the research governance training courses already undertaken as preparation for the project. These forms were started on 6th February 2017. These 'approved researcher' forms had to be approved by the institutional guarantor, ensuring that the research institution supports the researchers and adopts the responsibility for any misconduct, prior to being submitted to the ADRN proper. Institutional approval was granted on 10th February 2017; final ADRN approval was granted that same day. Note that this stage is no longer required since the conclusion of the ADRN, reflecting changes to the approvals process since the projects were undertaken.

Stage 3: Public benefit and privacy panel approval (prep time: 4.5 months; processing time: 1.5 months)

To ensure compliance with data protection law, it is necessary to demonstrate both a legal gateway by which data can be provided by data controllers (see *Stage 4* below) and a public benefit resulting from the research. Stage 3 of the linkage process dealt with obtaining permissions for data linkage and use from the Public Benefit and Privacy Panel (PBPP), whose role it is to weigh-up the potential benefits arising from proposed research projects against the risk of breaches in privacy. The PBPP provides a single-point of application for permissions

regarding health administrative data, where previously approval was required from the Caldicott Guardian of each health board involved[12]. Notably, this process was only required for the two census studies, as assessment of the public benefit and privacy of the SLS studies was combined with the data access approval process (see *Stage 4*). This is one of the key ways in which the process required by the SLS projects differed from the two census projects. We began drafting a single PBPP application form on 26th August 2016 with a view to simply amending the existing permissions for linkage and use established by previous work with the SMS1947 cohort[12]. However, it became apparent during the ethics process (see *Stage 1*) that the new project necessitated new approvals, and so two separate PBPP forms were drafted, one for each of the census studies. Note that two forms were required as the two census project addressed different research questions, despite the similarity in the datasets to be linked. We began these drafts on 21st November 2016. Initial drafts were submitted to the ADRC-S for feedback on 15th February 2017. Following extensive feedback from the ADRC-S support staff (each form was revised four times, from 23rd February to 9th April 2017), these two forms were submitted to the PBPP on 10th and 12th April 2017. Conditional approval was granted on 23rd May 2017, and the required amendments were re-submitted on 25th May 2017.

Stage 4: Data access approval (prep time: 2 months; processing time: 14 months)

After being granted ethical, network, and PBPP approval, the last approval to be obtained is that of the data controllers. Given that, to get to this stage, our studies had already been deemed ethical, feasible, legal, in the public interest, and reasonably secure, approval from the data controllers themselves might be considered to have been relatively trivial. In the case of some organisations this was, indeed, the case. The SLS studies were approved within 2 months of beginning the application process (see Box 1 and Figure 1). Again, this is thanks to the pre-linked nature of the SLS data and the unified approvals process. For the two census studies, NHS ISD – the organisation holding the majority of the health administrative data required by the studies – required no further approvals beyond those already obtained. The electronic Data Research and Innovation Service (eDRIS), acting as research coordinator for NHS ISD with regard to the requested health data, requested only proof of information governance training (i.e., certificates of completion). Approval from the individual acting as data controller for the Scottish Mental Survey 1947 was obtained on the same day as the relevant form was submitted – 13th March 2017.

Access to non-health administrative data, however, proved to be much more complicated. Four additional forms – a Privacy Impact Assessment and a Data Access application for each of the two census studies – were required by NRS and Scottish Government. Note that Privacy Impact Assessments have since been replaced by Data Protection Impact Assessments, and are now required for all new data linkage projects in the UK. We began drafting these forms on 11th April 2017 and sent them to the ADRC-S for initial review on 27th April 2017. After making changes according to the advice of ADRC-S support staff, finalised forms were sent to the ADRC-S on 16th June 2017. These forms, however, could not be submitted directly to the organisations and, instead, entered a queue. At the time, NRS and Scottish Government would only accept small 'batches' of around five projects at a time in order to avoid overloading their capacity; all projects within a batch would need to be processed and approved (or not) before the next batch would be accepted. As such, the ADRC-S retained a queue of batches ready for submission. Our two studies were part of the second batch, and so had to wait for the first batch to be examined and cleared before being submitted, let alone considered. The first batch was cleared on 7th July 2017, and the Data Access and Privacy Impact Assessment forms were submitted formally on 19th September 2017. On submission, these forms were distributed to the NRS Privacy Group, the Scottish Government Statistics PBPP and the Scottish Government lawyer for simultaneous assessment. On 4th December 2017 the presiding Scottish Government lawyer left the post, leading to a delay until the post could be filled. A new lawyer came into post in January, although this necessitated a reassessment of the forms by the new lawyer. While being considered by the Scottish Government legal team, Scottish Government Statistics PBPP approved the two census studies on 22nd January 2018. On 2nd March 2018, it became apparent that the new Scottish Government lawyer was unwilling to accept the legal gateway identified by census studies (Section 5 of the Census Act (Scotland)) or to approve the second batch based on the precedent of the first. Further investigation would be required to identify a new, more appropriate legal gateway for sharing census data. At this point, it was unclear how much time would be required for this investigation and how much of a delay would result. Due to the risk that census data would become available beyond the lifespan of the project, we decided to continue with the linkage between the other data sources for the two census studies. This necessitated an amendment to already-submitted PBPP forms (see Stage 2), which was submitted on 29th March 2018 and was approved on 3rd April 2018. However, a new legal gateway (Section 4 of the Census Act (Scotland)) was identified on 3rd April 2018, and the Scottish Government lawyer gave their approval for the two census projects on 17th April 2018.

The census projects were then passed to the Scottish Census Privacy Working Group, who review the privacy and security arrangements of studies. On 23rd May 2018, the Scottish Census Privacy Working Group asked for a revision of the intended census data retention period from 5 years (as per the eDRIS and National Safe Haven policies) to 2 years. An amendment to this effect was submitted on 25th May 2018, and access approval was gained in August 2018.

Stage 5: Data extraction and indexing (processing time: 10+ months)

After approval, data needs to be extracted and indexed before it is made available to the researcher. This process largely occurs 'behind the scenes', and is coordinated by the Trusted Third Party to help ensure privacy and minimise the transfer of personal data. Indexing – the process of assigning a random, unidentifiable index to each individual – was completed in

May 2019, several months after the end of project funding, and only for one of the census projects. Indexing delays have partly resulted from demand and staffing issues within the Trusted Third Party team (NRS Indexing). Although these indexes are now with data controllers for use in data extraction, no data can be transferred to the safe haven infrastructure until data sharing agreements are signed. These the legal contracts, which lay out the responsibilities of the organisations and researchers, are still being drafted and agreed between data controllers. As a result, the prospect of analysing data is still some way off. Meanwhile, access to SLS data was provided around 2 weeks after data access had been approved, and analysis was largely completed around 6 months after the data was made accessible.

Issues

Timing

One of the most important issues highlighted by the above narrative is the time taken to achieve non-health administrative data linkage (from 1st August 2016 to 7th June 2019 currently; see Figure 1). To date, the above project has taken 34 months, even before gaining access to the requested data. Stage 4, data access approval, has by far taken the most time, although it does not mark the end of the administrative process. The exception has been obtaining SLS data. As a standing database, the SLS has the advantage of well-established protocols and there being a single point of application. However, researchers may still face challenges when gaining access to and using SLS data (see Box 1). Furthermore, the restricted scope and relatively small sample size of the SLS may not be suitable for all researchers.

Previous efforts to obtain and link routinely-collected Scottish data for research purposes has been lengthy and complicated (e.g., 538 days)[12]. More recent changes in the Scottish data landscape, such as the Public Benefit and Privacy Panel, which replaces individual Caldicott Guardian approvals, should have improved the experience for researchers. This has generally been the case in regard to health administrative data, though data access is still prone to delays. For non-health data, whereas the number of forms to be submitted and organisations to be contacted has been reduced, the amount of time taken to obtain linked data has remained largely unchanged. Admittedly, the linkage project described here was much more ambitious than previous projects using the SMS1947 dataset and involved the linkage of more datasets from more data controllers. Notably, the current timescales are problematic for those conducting the research, particularly in academia where funding is time-limited. For the above project, data was not obtained before the end of the funding period and the post-doctoral researcher's contract. Taking the presented timescales as representative, the current situation essentially prohibits researchers, particularly early-stage researchers, from conducting projects involving linked non-health administrative data unless permissions are sought well in advance. For example, a full-time PhD student would have been required to submit their thesis within the time taken to obtain linked non-health administrative data.

Note that timing also affects those organisations set up to aid researchers in acquiring data. For example, the ADRC-S was funded for defined periods (1st October 2013 to 1st October 2018; by the ESRC), and was refunded but fundamentally re-specified within the lifetime of the described project (1st October 2018). At the same time, the funding for the ADRN was not renewed. These organisations were themselves preceded in Scotland by the Administrative Data Liaison Service and have since been superseded by the Administrative Data Research Partnership and the Scottish Centre for Administrative Data Research (SCADR). These organisations are judged on their success in obtaining new sources of administrative data, and on the number of research projects which are completed with their support. The changes to such organisations therefore reflects the challenges they face in producing results within their periods of funding, given the time taken to obtain data[14]. An unintended consequence of these changes has been the loss of much of the documentation which researchers use to learn about available datasets and necessary processes and which enable reproducible research[15]. In its lifetime, the ADRC-S supported over 70 projects, over 10 of which have now obtained data (linking over 25 distinct datasets) with over 40 projects still being sought under the new SCADR structure.

Process

The long timescales, particularly in the approval of data access requests, in part result from the relative infancy of the non-health administrative data landscape in Scotland. Whereas health administrative data controllers have developed clearer and more streamlined processes, non-health administrative data controllers are not yet at this stage. For those seeking only health administrative data, ethics, PBPP, indexing and extraction are required, though this process can still be lengthy[12]. However, where PBPP acts as a single point of health administrative data permissions, those seeking non-health administrative data must negotiate with and complete the governance procedures of all concerned data controllers, resulting in a complex and often unclear process.

The project described here has, to date, necessitated the submission of some 21 forms, including amendments (Table 1). For the most part, non-health administrative data controllers have been reflexively developing processes as data requests are submitted, and these processes have been prone to significant change. For example, the NRS and Scottish Government developed a process for dealing with census data requests in response to the first project submitted to them, a project examining end-of-life care (Schneider and Atherton, *In Preparation*) initially submitted in December 2015. As part of this new process, NRS and Scottish Government identified a legal gateway (Section 5 of the Census Act (Scotland)) necessary to allow them to share census data to researchers. Subsequent requests for census data followed this process, citing the same legal gateway. However, these projects were not provided with data due to a change in Scottish Government's legal interpretation regarding the appropriate legal gateway (see above). Such reflexive changes to policy and process by non-health administrative data controllers indicates the uncertainty with which they have taken to data sharing. Similar changes were seen during the development of health administrative data linkage procedures around 2013[12]. These changes also reflect a problematic culture within organisations in their perception of risk and public interest.

Previous reviews note that data controllers often design processes that disproportionately restrict data sharing in order to account for barriers to data sharing, whether real or perceived[16]. While it is important to get procedures right, unnecessary complexity and unexpected changes often lead to delays and damages the trust between researchers and data controllers. Trust requires appropriate, and not excessive, governance that can be well-understood by both sides of the process[17]. Whereas processes will continue to change as non-health administrative data controllers mature, the issue of trust may be partly addressed by their greater engagement early in the life of a research project. For example, data controllers could give conditional guarantees for data at the start of the permissions process, contingent on the research project acquiring ethical, public benefit and privacy approval. While this process would still need to respond to changing procedures, it would give some level of certainty for researchers and would ensure accountability for any subsequent delays or failures to provide data. However, appropriate incentives would be required to encourage such collaboration.

It is worth noting that some degree of complexity is necessary given the sensitivity and scale of the data being requested and the resulting risk to privacy and impact of improper use. The variety of data access procedures described here are in place to protect the privacy of individuals and their data, and it is important that researchers demonstrate their plans for and commitment to minimise any risks. Furthermore, data controllers such as NRS have responsibilities to care for the data under their charge which outweigh their responsibilities to share data for research. We therefore do not suggest that the data access process should be less thorough or strict. Indeed, streamlining the data access process (e.g., uniting all non-health data access approvals) should not come at the cost of an increased risk to data privacy nor at should it damage data controllers' responsibilities or reputation. However, improvements can be made to make the journey to obtaining data clearer and smoother. Reflecting on access to health administrative data, an early attempt to link SMS1947 records to routinely-collected health records found that the process took almost 2 years and required some 210 documents[12]. Since this time, the processes for accessing health administrative data have been streamlined through the development of standardised forms and the development of the PBPP as a single point of application for data access approvals in Scotland. While there can still be setbacks in accessing health administrative data – particularly for complex projects with multiple data sources – the overall result has been a faster process with fewer forms, as highlighted in the narrative presented here. Similar developments could benefit access to non-health administrative data as well as access to data in the rest of the UK.

Though necessary, the complexity of the permissions process (time, number of organisations, potential hurdles, etc.) creates a large barrier to entry for researchers. The current landscape necessitates a guide to identify the required points of application, to lay-out the process for each organisation, and to ensure applications are made in the most efficient order. Whereas research support officers in organisations such as the ADRC-S can (and do) help in this regard, this relies overly on the expert knowledge of specific individuals and on the existence (and capacity) of these support organisations. The future of non-health administrative research, then, is fragile: without clear and consistent processes, and without help to guide them between processes, new researchers would doubtless be lost.

Capacity

The relative infancy of non-health administrative data research is also reflected in the processing capacity of approvals panels and data controllers. Processing applications, indexing records and extracting data all require resources on the part of the organisation, both in terms of staffing and infrastructure. These resources are finite, and many of the organisations struggle to keep up with the rapidly increasing demand for non-health administrative data. Indeed, several organisations within the permissions and indexing process operate 'queues' for research projects. Many non-health administrative data controllers are expected to deal with data sharing requests using existing resources and funding, resulting in a reliance on staff who have other responsibilities beyond data sharing or on relatively small teams. In order to resolve the capacity problem and to speed up processing requests it is imperative that organisations commit more and dedicated staff and infrastructure, and that their funding enables these developments. This is not a new problem, and has been highlighted by previous initiatives in non-health administrative data[14]. Furthermore, this problem has already been recognised in regard to health data, and the capacity of health administrative data controllers is beginning to increase (e.g., "The research strategy for health and healthcare", Scottish Government, 2009); a similar effort needs to be made regarding non-health administrative data to help data controllers to deliver on their promises. Although recent investment has been made by the UK and Scottish governments in the form of the Administrative Data Partnership, this needs to be sufficiently targeted towards capacity-building and staffing to maximise the impact on data access.

A global perspective

Although the narrative presented describes the process of acquiring administrative data in Scotland, many of the experiences and challenges are common to other countries seeing an increase in administrative data research. Although Scotland is at the forefront in terms of the variety of health and non-health administrative data available to researchers, the process of obtaining data is largely the same in other countries. In England, for example, ethical and public benefit approvals are still needed before administrative data access requests will be considered and data extracted, although such approvals are sometimes regional and the Office for National Statistics and NHS Digital play a role in data coordination and linkage. The exception to this is perhaps in Scandinavian countries such as Sweden, in which health and non-health administrative data been utilised by researchers for much longer and data controllers are better provisioned for data requests.

Conclusion

With increasing interest in using non-health administrative data for research it is important to note the challenges that any such project might face. We hope that the narrative presented

above, detailing the journey of a project through various stages of the necessary processes in Scotland, helps to highlight these challenges. Big data approaches are powerful, and have the potential to be faster, capture larger more representative samples, and collect more varied types of data than other research methods, such as survey studies. However, it is important for those considering pursuing non-health administrative data to appreciate the time and effort required to eventually acquire data, if at all. Again, these challenges largely arise from the infancy of non-health administrative data organisations and their processes, relative to their counterparts in health data research. While still prone to data access delays, the development of health administrative data research should be somewhat of a model, with clearer processes and more investment in capacity making it easier to produce important and impactful research using big data. Large-scale investment in administrative data research, similar to recent investments in health data research (e.g., Health Data Research UK)[18], is only possible if the situation becomes more conducive to research. As these investments are starting to be made, such as the UK and Scottish governments' work in the Administrative Data Research Partnership, it is important to learn from previous efforts to make efficient use of resources. It is also important to incentivise organisations to participate fully in data sharing, and to encourage partnerships between data controllers and researchers. This may be achieved by the provision of funding and resources by research councils, by changing internal organisational goals, and by helping data controllers to benefit from research output[19]. As has been noted with health administrative data, there is a significant potential for harm should non-health administrative data not be shared and used[20]. As it stands, current attempts to obtain non-health administrative data are marked by uncertainty. Researchers are faced with a long journey through a complex and changeable landscape of permissions, approvals, and negotiations before reaching the prize of non-health administrative data. And non-health administrative data is quite the prize; with it, researchers have the potential to tackle the largest and most difficult problems faced by society.

Summary

What was already known on the topic:

- Following the success of research using routinely-collect health administrative data, there is increasing availability of routinely-collected non-health administrative data for research purposes.
- The process of acquiring non-health administrative data is less well-established than that for health administrative data, and is constantly changing.

What this study adds:

- A comprehensive and factual narrative detailing the steps required to gain access to linked health and non-health administrative data, from the perspective of a researcher.
- A review of the problems and barriers facing linked data research, as well as a discussion of potential solutions.

Data availability

No data are associated with this article.

Acknowledgements

An earlier version of this article can be found on Open Science Framework (DOI: https://doi.org/10.31219/osf.io/8tnfa).

References

1. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, *et al.*: **The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research.** *Eur J Epidemiol.* 2009; **24**(11): 659–667.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Shu X, Ji J, Li X, *et al.*: **Cancer risk in patients hospitalised for Graves' disease: a population-based cohort study in Sweden.** *Br J Cancer.* 2010; **102**(9): 1397–1399.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Li X, Sundquist J, Sundquist K: **Age-specific familial risks of psychotic disorders and schizophrenia: a nation-wide epidemiological study from Sweden.** *Schizophr Res.* 2007; **97**(1–3): 43–50.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Moster D, Lie RT, Markestad T: **Long-term medical and social consequences of preterm birth.** *N Engl J Med.* 2008; **359**(3): 262–273.
PubMed Abstract | Publisher Full Text
5. Foverskov E, Mortensen EL, Holm A, *et al.*: **Socioeconomic Position Across the Life Course and Cognitive Ability Later in Life: The Importance of Considering Early Cognitive Ability.** *J Aging Health.* 2017; 1–20.
6. Bjorklund A, Jantti M, Solon G: **Influences of nature and nurture on earnings variation: a report on a study of various sibling types in Sweden**. In: *Unequal Chances: Family Background and Economic Success.* (eds. S Bowles, H Gintis, M Osborne Groves), Princeton, NJ: Princeton Univ. Press. 2005; 145–164.
Reference Source
7. Deary IJ, Brett CE: **Predicting and retrodicting intelligence between childhood and old age in the 6-Day Sample of the Scottish Mental Survey 1947.** *Intelligence.* 2015; **50**: 1–9.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Batty GD, Calvin CM, Brett CE, *et al.*: **Childhood Body Weight in Relation to Cause-Specific Mortality: 67 Year Follow-up of Participants in the 1947 Scottish Mental Survey.** *Medicine (Baltimore).* 2016; **95**(6): e2263.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Calvin C, Batty GD, Der G, *et al.*: **Childhood intelligence in relation to major causes of death in 68 year follow-up: prospective population study.** *BMJ.* 2017; **357**: j2708.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Čukić I, Brett CE, Calvin CM, *et al.*: **Childhood IQ and survival to 79: Follow-up of 94% of the Scottish Mental Survey 1947.** *Intelligence.* 2017; **63**: 45–50.
PubMed Abstract | Publisher Full Text | Free Full Text
11. The Scottish Council for Research in Education: **The trend of Scottish intelligence**. London: University of London Press. 1949.
Reference Source
12. Brett CE, Deary IJ: **Realising health data linkage from a researcher's perspective: Following up the 6-Day Sample of the Scottish Mental Survey 1947.** *Longit Life Course Stud.* 2014; **5**(3): 283–298.
Publisher Full Text
13. RStudio Team: **RStudio: Integrated Development for R**. RStudio, Inc., Boston, MA. 2015.
14. Elias P: **The UK Administrative Data Research Network: Its genesis, progress,**

and future. *The Annals of the American Academy of Political and Social Science.* 2018; **675**(1): 184–201.
Publisher Full Text

15. Playford CJ, Gayle V, Connelly R, *et al.*: **Administrative social science data: The challenge of reproducible research.** *Big Data Soc.* 2016; **3**(2): 1–13.
Publisher Full Text

16. Laurie G, Stevens L: **Developing a public interest mandate for the governance and use of administrative data in the United Kingdom.** *J Law Soc.* 2016; **43**(3): 360–392.
Publisher Full Text

17. Sexton A, Shepherd E, Duke-Williams O, *et al.*: **A balance of trust in the use of government administrative data.** *Archival Science.* 2017; **17**(4): 305–330.
Publisher Full Text

18. Medical Research Council: **Director appointed for new UK health and biomedical informatics research institute**. London, UK: MRC, 2017; [Accessed 14 June 2018].
Reference Source

19. Card D, Chetty R, Feldstein MS, *et al.*: **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.* 2010.
Publisher Full Text

20. Jones KH, Laurie G, Stevens L, *et al.*: **The other side of the coin: Harm due to the non-use of health-related data.** *Int J Med Inform.* 2017; **97**: 43–51.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 29 November 2019

https://doi.org/10.21956/wellcomeopenres.17004.r36824

✔ **Michael Fleming**
Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK

I think the latest version of the manuscript is much improved and I am now happy that the authors have made efforts to address all previous comments where possible. I think the article reads more clearly now as a result.

It is now clear to the reader that these issues described relate to non-health data and, whilst similar issues certainly still exist when linking health data, the process of linking health data is more straightforward and quicker by comparison.

I think the paper highlights some very important issues and nicely summarizes some of the common problems when linking non-health data, whether it be trust or simply capacity issues. This clearly has an impact on research and the authors are correct to highlight the particularly detrimental impact on researchers on short term contracts, particularly PhD students. The authors also nicely summarize their arguments to have a more overarching and centralized approval process in place similar to PBPP for health data. This would be extremely beneficial. I just have two typographical errors to flag up.

1. Typo on page 4 - "the organisations involved, their role and they type...." they should be the.

2. Typo on page 8 - "Indeed streamlining the data process.....to data privacy nor at should it....." at should be deleted.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am a statistician and a research fellow who works in public health. Long standing interests include using novel record linkage techniques and statistical methods to analyse complex linked data, including routine administrative data, for research purposes across the spectrum of public health. My current research focuses on linkage of routine administrative health and non-health data to investigate educational and health outcomes related to childhood chronic conditions, early life factors, neonatal and childhood morbidity and maternal/obstetric factors.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 21 October 2019

https://doi.org/10.21956/wellcomeopenres.17004.r36823

Chris Playford
University of Exeter, Exeter, UK

I have no further comments to make. The revised article is much improved.

Competing Interests: I know Dr Matthew Iveson and Prof Ian Deary from when I used to work at the Administrative Data Research Centre – Scotland (University of Edinburgh) between 2014 and 2017. We have not co-authored any papers together or collaborated directly on a piece of work. I do not believe this has affected my objectivity when reviewing this manuscript.

Reviewer Expertise: I am a sociologist working in the fields of social stratification and the sociology of education. My work has focused on modelling the role of family background on educational attainment with a substantive interest in inequality and disadvantage. I specialise in the secondary analysis of large-scale survey and administrative data.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 28 August 2019

https://doi.org/10.21956/wellcomeopenres.16744.r36032

Michael Fleming
Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK

This is a useful personal account of a researcher's struggles to gain access to, and link together, routine administrative datasets in Scotland and is a useful reference for other researchers aiming to do the same. It highlights a real issue which needs addressing and it is one which I have personally encountered.

However, I feel that the article could be more general and broader in scope, with more developed conclusions and more discussion of some of the counter arguments for having such strict data governance in place. I have several comments which I feel would improve this article.

1. Throughout the manuscript the authors describe health data as ‘routinely collected health data’ and non-health data as ‘administrative data’. I do not like this terminology because healthcare data is also administrative data and is described this way throughout the literature. Therefore, I think the authors should stick to describing healthcare data as ‘routine administrative healthcare data’ and non-health data as ‘routine administrative non-health data’. We can of course get health and non-health data which isn’t administrative or routinely collected but these are not the focus of this manuscript.

2. The authors compound the problem by stating on more than one occasion that ‘researchers have recently extended their sights to routinely collected administrative data’ or that ‘access to administrative data is a relatively recent development’. These statements are not correct because healthcare data is (correctly) commonly thought of as routinely collected administrative data and has been accessed for decades. Be specific and highlight that you are referring to non-health data. Also, some context here would be helpful for the non-health data. What do the authors mean by more recently? Within the last 20, 10, or 5 years?

3. The researcher’s perspective and narrative detailing the journey is very detailed and could, I feel, be greatly reduced whilst still getting the same message across. At present I feel this is too lengthy and dominates over the discussion of the wider issues and contexts which are covered in comparatively less in depth, but which are much more pertinent to the wider research community.

4. Following on from the last point, I feel that the title of the manuscript should therefore be reworded as it is a little misleading. The manuscript currently describes a particular set of circumstances arising from a research project aiming to analyse a particular set of administrative data sets so this should be specified and reflected more in the title. Alternatively, if the title remains the same then I think more focus needs to be put on the wider, more general, issues and less on the detailed specifics of the researchers own project narrative. I expected a more general review paper when I saw the title.

5. I agree with reviewer 1 that the authors need to give some more detail around SLS and highlight that this is a longstanding data resource which has been linked many times in the past hence accounting for the quicker access.

6. I agree with reviewer 1 that the authors should highlight that these difficulties have an important and detrimental knock-on effect particularly for PhD students who can’t gain access within the allotted study time, or indeed for researchers on short term contracts e.g. fellowships.

7. I think the authors need to provide more in the way of counter arguments. For example, why is governance around routine administrative data so strict and what are some of the drawbacks of making the approval processes easier? What are the authors own thoughts around the reasons for the current landscape being the way it is in terms of risks around accessing data? These processes are in place for a reason so it would be interesting to see some acknowledgement of these along with references to make this a more balanced account. Also, some more discussion around the background both in terms of the current processes for health (e.g. PBPP) and non-health data would be really useful to help navigate the reader and provide more general guidance around what the most important steps are when trying to apply to access health and

non-health administrative datasets. I think the scope of the manuscript is too narrow and specific to the datasets within the researcher's particular study and is not general enough given its very general and over-arching title.

8. These issues are still not just peculiar to non-health administrative data. Indeed, depending on the complexity of the project and the datasets involved, the approvals process to access administrative health data can still be much lengthier than it ought to be. I think this needs to be acknowledged within the manuscript. In terms of solutions for non-health data (and indeed health data), the authors mainly focus on adding capacity and improving large-scale investment. I'd like to have seen these ideas discussed in slightly more depth.

Is the rationale for the Open Letter provided in sufficient detail?
Yes

Does the article adequately reference differing views and opinions?
No

Are all factual statements correct, and are statements and arguments made adequately supported by citations?
No

Is the Open Letter written in accessible language?
Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am a statistician and a research fellow who works in public health. Long standing interests include using novel record linkage techniques and statistical methods to analyse complex linked data, including routine administrative data, for research purposes across the spectrum of public health. My current research focuses on linkage of routine administrative health and non-health data to investigate educational and health outcomes related to childhood chronic conditions, early life factors, neonatal and childhood morbidity and maternal/obstetric factors.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Oct 2019

Matthew Iveson, The University of Edinburgh, Edinburgh, UK

Dear Dr Fleming,

We very much appreciate the time and care that you have taken to review our manuscript. Your comments have helped to clarify several points and to develop the discussion of data access issues. We believe that the manuscript is more comprehensive and useful as a result of these

changes. We have submitted an updated manuscript incorporating the comments of both reviewers. However, we would like to take the opportunity to outline our response to your individual comments below.

Response to comments

This is a useful personal account of a researcher's struggles to gain access to, and link together, routine administrative datasets in Scotland and is a useful reference for other researchers aiming to do the same. It highlights a real issue which needs addressing and it is one which I have personally encountered. However, I feel that the article could be more general and broader in scope, with more developed conclusions and more discussion of some of the counter arguments for having such strict data governance in place. I have several comments which I feel would improve this article.

Response: We appreciate the reviewer's concerns. From the start we have aimed this manuscript as an informative account of one specific set of projects, and we have acknowledged that this may not represent everyone's experiences with routinely-collected data. Accurately portraying this journey necessitates a level of detail that we hope serves to motivate discussion about steps in the process which everyone undergoes. This being said, we agree with the reviewer and we have now attempted to broaden the discussion of the administrative data landscape and to reflect on changes to the system over time as well as organisational attitudes to data sharing (e.g., Process section, Page 11).
It is also not our intention to suggest that these processes should not be in place or that they should be less strict. Instead, we suggest that they are often disproportionate and fragile, being subject to changes and delays. We also do not suggest changing these processes necessarily, but instead advocate for better communication and collaboration between data controllers and researchers so as to manage expectations going in. As suggested by a later comment, we have added a paragraph to acknowledge that some complexity in the process is necessary to safeguard data (Process section, Page 11).

Throughout the manuscript the authors describe health data as 'routinely collected health data' and non-health data as 'administrative data'. I do not like this terminology because healthcare data is also administrative data and is described this way throughout the literature. Therefore, I think the authors should stick to describing healthcare data as 'routine administrative healthcare data' and non-health data as 'routine administrative non-health data'. We can of course get health and non-health data which isn't administrative or routinely collected but these are not the focus of this manuscript.
The authors compound the problem by stating on more than one occasion that 'researchers have recently extended their sights to routinely collected administrative data' or that 'access to administrative data is a relatively recent development'. These statements are not correct because healthcare data is (correctly) commonly thought of as routinely collected administrative data and has been accessed for decades. Be specific and highlight that you are referring to non-health data. Also, some context here would be helpful for the non-health data. What do the authors mean by more recently? Within the last 20, 10, or 5 years?

Response: We have amended our terms as the reviewer suggests. We made our original distinctions due to the focus of organisations such as the Administrative Data Research Network on obtaining non-health routinely-collected data. However, we appreciate that this may have caused some confusion. We have also clarified that much of the

development has been regarding non-health administrative data. We have also replaced "recently" with "In the last decade" (Page 3).

The researcher's perspective and narrative detailing the journey is very detailed and could, I feel, be greatly reduced whilst still getting the same message across. At present I feel this is too lengthy and dominates over the discussion of the wider issues and contexts which are covered in comparatively less in depth, but which are much more pertinent to the wider research community.

Response: The manuscript is intended as an account of a researcher's journey to accessing administrative data. Its primary aim is to inform the reader – particularly new researchers – about the process and challenges, and therefore the level of detail is necessary. Whereas we do discuss the implications and the relevance to the wider data linkage scene, we agree that this was not sufficiently deep. However, the issues of ethics and governance have been discussed in greater detail elsewhere, and we have made sure to point the reader to them as well as raising the issues (e.g., Playford *et al*., 2016; Elias, 2018). As suggested, we have developed our discussion of the timing, process and capacity issues while referencing these other papers, while acknowledging that these are not new problems (Pages 10-12).

Following on from the last point, I feel that the title of the manuscript should therefore be reworded as it is a little misleading. The manuscript currently describes a particular set of circumstances arising from a research project aiming to analyse a particular set of administrative data sets so this should be specified and reflected more in the title. Alternatively, if the title remains the same then I think more focus needs to be put on the wider, more general, issues and less on the detailed specifics of the researchers own project narrative. I expected a more general review paper when I saw the title.

Response: We have amended the title to clarify that the manuscript is predominantly a narrative from a researcher's perspective and relates primarily to non-health administrative data.

I agree with reviewer 1 that the authors need to give some more detail around SLS and highlight that this is a longstanding data resource which has been linked many times in the past hence accounting for the quicker access.

Response: As suggested by both reviewers, we have added an explanation of the SLS as a standing pre-linked resource where it is first introduced (Page 4). We have also added that SLS approvals and extraction are typically faster than non-SLS projects, referring to Figure 1.

I agree with reviewer 1 that the authors should highlight that these difficulties have an important and detrimental knock-on effect particularly for PhD students who can't gain access within the allotted study time, or indeed for researchers on short term contracts e.g. fellowships.

Response: We have highlighted the particular issue for PhD and early-career researchers in the second paragraph on Timing (page 10).

I think the authors need to provide more in the way of counter arguments. For example, why is governance around routine administrative data so strict and what are some of the drawbacks of

making the approval processes easier? What are the authors own thoughts around the reasons for the current landscape being the way it is in terms of risks around accessing data? These processes are in place for a reason so it would be interesting to see some acknowledgement of these along with references to make this a more balanced account. Also, some more discussion around the background both in terms of the current processes for health (e.g. PBPP) and non-health data would be really useful to help navigate the reader and provide more general guidance around what the most important steps are when trying to apply to access health and non-health administrative datasets. I think the scope of the manuscript is too narrow and specific to the datasets within the researcher's particular study and is not general enough given its very general and over-arching title.

Response: We would like to thank the reviewer for highlighting the need for a balanced discussion regarding the information governance process. We agree that the process is complex for a reason and have incorporated these suggestions into the manuscript. In particular, we have added the following to the Process section (Page 11):

- "It is worth noting that some degree of complexity is necessary given the sensitivity and scale of the data being requested and the resulting risk to privacy and impact of improper use. The variety of data access procedures described here are in place to protect the privacy of individuals and their data, and it is important that researchers demonstrate their plans for and commitment to minimise any risks. Furthermore data controllers such as NRS have responsibilities to care for the data under their charge which outweigh their responsibilities to share data for research. We therefore do not suggest that the data access process should be less thorough or strict. Indeed, streamlining the data access process (e.g., uniting all non-health data access approvals) should not come at the cost of an increased risk to data privacy nor at should it damage data controllers' responsibilities or reputation. However, improvements can be made to make the journey to obtaining data clearer and smoother."

Notably, we do not suggest that the governance and access process should become less strict or easier. Instead, we suggest that the process could be smoother and more streamlined, and that timescales could be reduced. This requires further investment in capacity and better communication between researchers and data controllers. A deeper discussion of the governance process is provided by other papers (e.g., Playford *et al*., 2016; Elias, 2018), and we have noted this throughout the Issues section (Pages 10-12). More detail about the current processes is included in the sections about Stage 1-5. For example, the background to the PBPP process is described in Stage 3 and in the Process section. We have further clarified and distinguished the steps needed to access health and non-health data in the Process section (Page 11).

These issues are still not just peculiar to non-health administrative data. Indeed, depending on the complexity of the project and the datasets involved, the approvals process to access administrative health data can still be much lengthier than it ought to be. I think this needs to be acknowledged within the manuscript. In terms of solutions for non-health data (and indeed health data), the authors mainly focus on adding capacity and improving large-scale investment. I'd like to have seen these ideas discussed in slightly more depth.

Response: The issues of complexity and timing have been raised in regards to health-related administrative data elsewhere. For example, an early attempt to link the same cohort to routinely-collected health data across the UK found that the process (pre-PBPP) took almost 2 years and required some 210 documents (Brett & Deary, 2014).

Since this paper the process for accessing health data has been streamlined (e.g., the introduction of the PBPP process) and has gotten faster for the majority of users. We have highlighted the development in health data access (and the previous paper) in the Process section (Pages 11-12), in particular noting that access to health data has become streamlined if not faster. We have also acknowledged that this streamlining is not perfect, and that health data access can still be longer than is necessary in both the Timing section (Page 10) and the Process section (Page 12).
We have discussed the potential solutions to data access problems – capacity, investment and communication – in more detail in the relevant sections and in the Conclusion (Page 13).

Competing Interests: No competing interests were disclosed.

Reviewer Report 21 June 2019

https://doi.org/10.21956/wellcomeopenres.16744.r35782

Chris Playford
University of Exeter, Exeter, UK

This is a helpful article that serves to illustrate the current challenges for researchers working with administrative data in Scotland. I recommend this article be approved for indexing because the account provides a useful contribution to understanding the current governance of administrative data access and the impact this can have on academic research. I have a number of recommendations for how the article might be improved. The first are minor points about clarifying the existing text. The final part of this review provides suggestion of how the scope of the conclusions might be developed further for the benefit of researchers, funders, and those working with administrative data in the UK. This is important for improving access to administrative data for future researchers and recognising the broader issues in which this process is currently contextualised. This final set of recommendations help clarify my response to the question: *“Does the article adequately reference differing views and opinions?”*

Firstly, here are my specific points of clarification:

On page 3, you mentioned the Scottish Longitudinal Study (SLS). It would be helpful to provide more context for readers unfamiliar with the SLS (see https://sls.lscs.ac.uk/about/). For instance, as you are aware, the SLS is a “standing resource” whereas the other linked data require that the data are linked for the first time. You mention this on page 6 but it would be useful to know earlier in the article. This explains in part why the data extraction and indexing differ so greatly, as you show helpfully in Figure 1.

Table 1 has a number of blank cells where I would have expected numeric figures to be shown – this needs rectifying.

Please be consistent when describing dates – I would recommend you always include the year otherwise

it is easy to lose track of the general timescale. For example, on page 6 you report a number of dates without the year listed.

Box 1 on page 5 is helpful as a means of comparing the process when using the SLS data. It is helpful to reiterate to the reader frequently that the SLS is quite different to the other administrative data projects which have not been previously linked.

On page 7, you refer to the funding timescale of the ADRC-S and ADRN. I would suggest in your conclusion that you add that that these were preceded in Scotland by the Administrative Data Liaison Service (ADLS) and have been superseded by the Administrative Data Research Partnership (ADRP). It is useful for the reader to be aware of this evolution. This is relevant so that those working in the field can see the need to learn from previous initiatives. As a broader point, a further challenge to researchers in Scotland and the UK is that much of the documented activity and resources these projects created does not appear to be available as much of the web content has either been removed. This is an impediment to current and future researchers seeking to learn about potential administrative datasets, processes of accessing data and the reproducibility of administrative data research (see Playford *et al*., (2016)[1]).

Here are some webpages that may be helpful.

Administrative Data Liaison Service (ADLS)
https://www.spi.ox.ac.uk/administrative-data-liaison-service-2014
https://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/elliot.pdf

Administrative Data Research Network (ADRN)
https://esrc.ukri.org/research/our-research/administrative-data-research-network/

For a general overview of the ADRN and lessons learned, see Elias (2018)[2].

New Administrative Data Research Partnership (ADRP)
https://esrc.ukri.org/news-events-and-publications/news/news-items/addressing-major-societal-challenge

Scottish ADRP
http://www.epcc.ed.ac.uk/blog/2018/scottish-administrative-data-research-partnership

On page 7, when you describe the issues faced, I would suggest you add that the timescales involved in this project would have precluded a PhD student from using linked administrative data as part of their thesis.

You make a number of important points when reflecting on the number of forms that were requested and the changes to the process that were developed during the project.

My final points relate to developing your conclusions further. These reflect briefly on the implications for researchers wishing to use administrative data. It would be helpful for the reader to understand the administrative data context in Scotland better. The following points include some references you may find useful.

Initiatives such as the recent Administrative Data Research Partnership (ADRP) indicate the substantial investment by the UK government into the use of administrative data for research purposes in the social

sciences. There is a clear desire to use this money wisely and efficiently. To do so, I would suggest that we must learn from previous efforts (see also Elias (2018)[2]). Your paper is a practical exemplar of the challenges faced by researchers working in the field. It is more broadly recognised that accessing administrative data is currently tricky and time-consuming (Connelly *et al*., (2016)[3], Harron *et al*., (2017)[4]).

I would encourage you to reflect in your conclusions on the wider organisational context in which administrative data research occurs. For example, although initially focusing on the legal gateways through which data could be accessed, Laurie and Stevens (2016)[5] identified that problematic organizational culture (particularly the perception of risk) was a significant barrier to proportionate governance. Sexton *et al*. (2017, p.327)[6] argue that: *"In trustworthy systems and processes, a balance must be struck between appropriate monitoring in the system whilst ensuring against excessive auditing that may counterproductively contribute to the erosion of trust."* Whilst describing administrative data access in the USA, Card *et al*. (2010)[7] provide some helpful suggestions of how to incentivise administrative data access and output for agencies involved. I would encourage you to reflect further on these points in your conclusions. Finally, there is a risk of potential harm due to the non-use of data (Jones *et al*., (2017)[8]). It is therefore important that the barriers that you describe are overcome if others are to benefit from this work.

References

1. Playford C, Gayle V, Connelly R, Gray A: Administrative social science data: The challenge of reproducible research. *Big Data & Society*. 2016; **3** (2). Publisher Full Text
2. Elias P: The UK Administrative Data Research Network: Its Genesis, Progress, and Future. *The ANNALS of the American Academy of Political and Social Science*. 2018; **675** (1): 184-201 Publisher Full Text
3. Connelly R, Playford CJ, Gayle V, Dibben C: The role of administrative data in the big data revolution in social science research.*Soc Sci Res*. **59**: 1-12 PubMed Abstract | Publisher Full Text
4. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, Goldstein H: Challenges in administrative data linkage for research.*Big Data Soc*. 2017; **4** (2): 2053951717745678 PubMed Abstract | Publisher Full Text
5. Laurie G, Stevens L: Developing a Public Interest Mandate for the Governance and Use of Administrative Data in the United Kingdom. *Journal of Law and Society*. 2016; **43** (3): 360-392 Publisher Full Text
6. Sexton A, Shepherd E, Duke-Williams O, Eveleigh A: A balance of trust in the use of government administrative data. *Archival Science*. 2017; **17** (4): 305-330 Publisher Full Text
7. Card D, Chetty R, Feldstein M, Saez E: Expanding Access to Administrative Data for Research in the United States. *SSRN Electronic Journal*. 2010. Publisher Full Text
8. Jones KH, Laurie G, Stevens L, Dobbs C, Ford DV, Lea N: The other side of the coin: Harm due to the non-use of health-related data.*Int J Med Inform*. **97**: 43-51 PubMed Abstract | Publisher Full Text

Is the rationale for the Open Letter provided in sufficient detail?

Yes

Does the article adequately reference differing views and opinions?

Partly

Are all factual statements correct, and are statements and arguments made adequately supported by citations?

Yes

Is the Open Letter written in accessible language?
Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?
Yes

Competing Interests: I know Matthew Iveson and Ian Deary from when I used to work at the Administrative Data Research Centre – Scotland (University of Edinburgh) between 2014 and 2017. We have not co-authored any papers together or collaborated directly on a piece of work. I do not believe this has affected my objectivity when reviewing this manuscript.

Reviewer Expertise: I am a sociologist working in the fields of social stratification and the sociology of education. My work has focused on modelling the role of family background on educational attainment with a substantive interest in inequality and disadvantage. I specialise in the secondary analysis of large-scale survey and administrative data.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 18 Oct 2019

Matthew Iveson, The University of Edinburgh, Edinburgh, UK

Dear Dr Playford,

We very much appreciate the time and care that you have taken to review our manuscript. Your comments have helped to clarify several points and to develop the discussion of data access issues. We believe that the manuscript is more comprehensive and useful as a result of these changes. We have submitted an updated manuscript incorporating the comments of both reviewers. However, we would like to take the opportunity to outline our response to your individual comments below.

Response to comments
This is a helpful article that serves to illustrate the current challenges for researchers working with administrative data in Scotland. I recommend this article be approved for indexing because the account provides a useful contribution to understanding the current governance of administrative data access and the impact this can have on academic research. I have a number of recommendations for how the article might be improved. The first are minor points about clarifying the existing text. The final part of this review provides suggestion of how the scope of the conclusions might be developed further for the benefit of researchers, funders, and those working with administrative data in the UK. This is important for improving access to administrative data for future researchers and recognising the broader issues in which this process is currently contextualised. This final set of recommendations help clarify my response to the question: “Does the article adequately reference differing views and opinions?”
Firstly, here are my specific points of clarification:

- On page 3, you mentioned the Scottish Longitudinal Study (SLS). It would be helpful to provide more context for readers unfamiliar with the SLS (see https://sls.lscs.ac.uk/about/). For instance, as you are aware, the SLS is a “standing resource” whereas the other linked data require that the data are linked for the first time. You mention this on page 6 but it would

be useful to know earlier in the article. This explains in part why the data extraction and indexing differ so greatly, as you show helpfully in Figure 1.

Response: As suggested, we have added explanation of the SLS as a standing pre-linked resource where it is first introduced (Page 4). We have also added that SLS approvals and extraction are typically faster than non-SLS projects, referring to Figure 1.

- Table 1 has a number of blank cells where I would have expected numeric figures to be shown – this needs rectifying.

Response: The blank columns indicated where an organisation did not have a 'data controller' row, and therefore did not have an applicable data type. We have amended Table 1 to remove the 'Type of data' column and instead expand the 'Role' column to reference the type of data for a given data controller.

- Please be consistent when describing dates – I would recommend you always include the year otherwise it is easy to lose track of the general timescale. For example, on page 6 you report a number of dates without the year listed.

Response: We have added the year to all dates, as suggested.

- Box 1 on page 5 is helpful as a means of comparing the process when using the SLS data. It is helpful to reiterate to the reader frequently that the SLS is quite different to the other administrative data projects which have not been previously linked.

Response: We have strengthened the description of the SLS in Box 1 to highlight the differences in procedure/requirements between the SLS and the two 'unlinked' census projects. We have also added sentences to Stage 3 (page 6) and Stage 4 (page 7) to better highlight that the SLS process is markedly different for these stages in particular.

- On page 7, you refer to the funding timescale of the ADRC-S and ADRN. I would suggest in your conclusion that you add that that these were preceded in Scotland by the Administrative Data Liaison Service (ADLS) and have been superseded by the Administrative Data Research Partnership (ADRP). It is useful for the reader to be aware of this evolution. This is relevant so that those working in the field can see the need to learn from previous initiatives. As a broader point, a further challenge to researchers in Scotland and the UK is that much of the documented activity and resources these projects created does not appear to be available as much of the web content has either been removed. This is an impediment to current and future researchers seeking to learn about potential administrative datasets, processes of accessing data and the reproducibility of administrative data research (see Playford et al., (2016)[1]).

Here are some webpages that may be helpful.
Administrative Data Liaison Service (ADLS)
https://www.spi.ox.ac.uk/administrative-data-liaison-service-2014
https://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/elliot.pdf
Administrative Data Research Network (ADRN)
https://esrc.ukri.org/research/our-research/administrative-data-research-network/
For a general overview of the ADRN and lessons learned, see Elias (2018)[2].
New Administrative Data Research Partnership (ADRP)
https://esrc.ukri.org/news-events-and-publications/news/news-items/addressing-major-societal-challer

Scottish ADRP
http://www.epcc.ed.ac.uk/blog/2018/scottish-administrative-data-research-partnership

Response: We would like to thank the reviewer for their suggestions and for the links to the relevant organisations. We have added detail of the preceding and succeeding organisations to the Timing section where we reference the funding timescale of the

ADRC-S (page 10) and have highlighted the potential loss of important documentation. We have also referenced Elias (2018) and Playford *et al*. (2016) in our attempts to better contextualise the ADRC-S and its evolution.

On page 7, when you describe the issues faced, I would suggest you add that the timescales involved in this project would have precluded a PhD student from using linked administrative data as part of their thesis.

Response: We have noted this in the second paragraph on Timing (page 10).

You make a number of important points when reflecting on the number of forms that were requested and the changes to the process that were developed during the project.

Response: We thank the reviewer for the comment, and hope that presenting simple statistics such as the number of forms required helps researchers to manage their expectations going into administrative data research.

My final points relate to developing your conclusions further. These reflect briefly on the implications for researchers wishing to use administrative data. It would be helpful for the reader to understand the administrative data context in Scotland better. The following points include some references you may find useful.
Initiatives such as the recent Administrative Data Research Partnership (ADRP) indicate the substantial investment by the UK government into the use of administrative data for research purposes in the social sciences. There is a clear desire to use this money wisely and efficiently. To do so, I would suggest that we must learn from previous efforts (see also Elias (2018)[2]). Your paper is a practical exemplar of the challenges faced by researchers working in the field. It is more broadly recognised that accessing administrative data is currently tricky and time-consuming (Connelly et al., (2016)[3], Harron et al., (2017)[4]).

Response: We have added reference to the significant investment by the UK government, the drive for efficiency, and the need to learn lessons to the Conclusion (Page 13). We have also referenced this issue in the Capacity section (Pages 12-13), as any investment needs to be sufficiently targeted for efficiency.

I would encourage you to reflect in your conclusions on the wider organisational context in which administrative data research occurs. For example, although initially focusing on the legal gateways through which data could be accessed, Laurie and Stevens (2016)[5] identified that problematic organizational culture (particularly the perception of risk) was a significant barrier to proportionate governance. Sexton et al. (2017, p.327)[6] argue that: “In trustworthy systems and processes, a balance must be struck between appropriate monitoring in the system whilst ensuring against excessive auditing that may counterproductively contribute to the erosion of trust.” Whilst describing administrative data access in the USA, Card et al. (2010)[7] provide some helpful suggestions of how to incentivise administrative data access and output for agencies involved. I would encourage you to reflect further on these points in your conclusions. Finally, there is a risk of potential harm due to the non-use of data (Jones et al., (2017)[8]). It is therefore important that the barriers that you describe are overcome if others are to benefit from this work.

Response: We have added discussion of the organisational context, including the need

for a balanced governance procedure, in the Process section (Page 11). We have also added a note regarding incentives and the consequences of data non-use in the Conclusion (Page 13). We would also like to thank the reviewer for the suggested references, and we have included them in the relevant sections.

Competing Interests: Dr Playford and I were both affiliated with the Administrative Data Research Centre Scotland. However, we did not co-author any papers or collaborate on any projects. I do not believe that this has influenced my of the peer review report.