OPEN LETTER

Navigating the landscape of administrative data in Scotland

[version 1; peer review: 1 approved]

Matthew H. Iveson, Ian J. Deary

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

Abstract

Background: There is growing interest in using routinely collected data for research purposes. Following the success of research using routinely collected healthcare data, attention has turned to leveraging administrative 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 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 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
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.

Copyright: © 2019 Iveson MH and Deary IJ. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Iveson MH and Deary IJ. Navigating the landscape of administrative data in Scotland [version 1; peer review: 1 approved] Wellcome Open Research 2019, 4:97 (https://doi.org/10.12688/wellcomeopenres.15336.1)

First published: 17 Jun 2019, 4:97 (https://doi.org/10.12688/wellcomeopenres.15336.1)
**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. These studies have significantly improved our understanding of issues such as cancer, mental health conditions, pre-term birth, cognitive ageing, socioeconomic inequality, etc. In Scotland, previous work has already leveraged routinely collected health 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.

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). More recently, 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 administrative data on their behalf. Furthermore, many of the organisations controlling 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 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) to routinely collected administrative data, including the Scottish Census. This follows on from and extends similar efforts to link the same cohort to routinely-collected health data, carried out before major changes in the Scottish landscape of big data.

While previous efforts have used linked SMS1947 and health data to investigate life-course determinants of cause-specific mortality, the current project sought to extend this linkage to 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 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 Scottish Longitudinal Study (SLS). The 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 organisations involved, their role, and the type of data they hold are summarised in Table 1. We provide this narrative in the hopes that it will inform researchers who are considering working with 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 anonymised health data linkage had been achieved for the SMS1947 cohort as part of a previous project, 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...
Table 1. Organisations, their role, and number of forms required. Merged cells indicate shared involvement.

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Role</th>
<th>Type of Data Held</th>
<th>Forms submitted</th>
<th>Amendments submitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative Data Research Centre - Scotland</td>
<td>Research Support</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACCORD</td>
<td>Clinical Sponsor</td>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>NHS Research Ethics Committee</td>
<td>Ethical Approval</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>University of Edinburgh Legal Services</td>
<td>Institutional Guarantor</td>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Administrative Data Research Network</td>
<td>Network Approval and Resources</td>
<td></td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Public Benefit and Privacy Panel</td>
<td>Public Benefit and Privacy Approval</td>
<td></td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>NHS Information Services Division</td>
<td>Data Controller</td>
<td>Health Data</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Electronic Data Research and Innovation Service</td>
<td>Research Coordinator for NHS ISD</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>National Records of Scotland</td>
<td>Data Controller/Trusted Third Party</td>
<td>Births and Deaths Data</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Scottish Government</td>
<td>Data Controller</td>
<td>Census Data</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Scottish Longitudinal Study</td>
<td>Data Controller</td>
<td>Various Administrative Data</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

linkage of health 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, and for final review on 18th November. 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.

Stage 3: Public benefit and privacy panel approval (prep time: 4.5 months; processing time: 1.5 months)
In order for linkages to be conducted and data to be used, 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 data, where previously approval was required from the Caldicott Guardian of each health board involved[1]. 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). 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. 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. 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). For the two census studies, NHS ISD – the organisation holding the majority of the health data required by the studies – required no further approvals beyond those already obtained. The electronic Data Research

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 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. 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 were conducted in R Studio using specialist packages. 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.
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 census 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. 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, and the Data Access and Privacy Impact Assessment forms were submitted formally on 19th September. 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 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, 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 has been transferred to the safe haven infrastructure as of yet, and 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 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)\(^2\). 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. 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 administrative data unless permissions are sought well in advance.

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 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. To date, the ADRC-S has supported over 70 projects, over 10 of which have now obtained data (linking over 25 distinct datasets) with over 40 projects still being sought.

Process
The long timescales, particularly in the approval of data access requests, in part result from the relative infancy of the administrative data landscape in Scotland. Whereas health data controllers have developed clearer and more streamlined processes (e.g., the development of the PBPP as a single point of health data permissions), administrative data controllers are not yet at this stage. The project described here has, to date, necessitated the submission of some 21 forms (including amendments; Table 1). For the most part, 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 administrative data controllers indicates the uncertainty with which they have taken to data sharing. While it is important to get procedures right, unexpected changes often lead to delays and damages the trust between researchers and data controllers. Whereas processes will continue to change as 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.

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 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 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 administrative data. Indeed, several organisations within the permissions and indexing process operate ‘queues’ for research projects. Many 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 need has already been recognised in regard to health data, and the capacity of health 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 administrative data to help data controllers to deliver on their promises.

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 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 data access requests will be considered and data extracted, although such approvals are sometimes regional and the Office for National Statistics plays a role in data coordination and linkage. The exception to this is perhaps in Scandinavian countries such as Sweden, in which administrative data been utilised by...
researchers for much longer and data controllers are better provisioned for data requests.

Conclusion
With increasing interest in using 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 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 administrative data organisations and their processes, relative to their counterparts in health data research. Indeed, the development of health 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), is only possible if the situation becomes more conducive to research. As it stands, current attempts to obtain 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 administrative data. And 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-collected healthcare data, there is increasing availability of routinely-collected administrative data for research purposes.

What this study adds:
- The process of acquiring administrative data is less well-established than that for healthcare data, and is constantly changing.

Data availability
No data are associated with this article.

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.

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

References

10. Čučiši I, Brett CE, Calvin CM, et al.: Childhood IQ and survival to 79: Follow-up of


Reference Source
Open Peer Review

Current Peer Review Status: ✓

Version 1

Reviewer Report 21 June 2019

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

© 2019 Playford C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 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)).

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).

New Administrative Data Research Partnership (ADRP)

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)). 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), Harron et al., (2017)).

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., 2017⁸). It is therefore important that the barriers that you describe are overcome if others are to benefit from this work.

References

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.