PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement
AUTHORS	Taylor, Julie; Crowe, Sonya; Espuny Pujol, Ferran; Franklin, Rodney; Feltbower, Richard; Norman, Lee; Doidge, James; Gould, Doug; Pagel, Christina

VERSION 1 – REVIEW

REVIEWER	OReilly, Dermot Queens University Belfast, Epidemiology and Public Health
REVIEW RETURNED	17-Apr-2021

GENERAL COMMENTS	Thank you for asking me to review this interesting and well written paper. I especially liked the title, though they may also want to consider…'the long and winding road' or 'Chasing the rainbow's end'… This is an unusual paper and the authors have done well to try and make it fit into the usual academic paper format. I entirely understand and appreciate the frustration and exasperation that they must have felt for much of this journey, and I think the picture will be well recognised by many in the field. Indeed, there are some of us who have been trying to access particular datasets from unnamed Departments for almost seven years. I have no real problems with the paper and think it should be published both as a potential road map for others to follow and a reassurance to others that their case is not unique; though it may deter some from embarking on the journey at all! I would ask the authors whether they thought their decision to ask individual data controllers to link their data to NCHDA identifiers was a food one and may have added to their woes and troubles? There are many important points here such as the need for understanding funders and the difficulties of identifying in advance how much time the various processes will take. I do not know the English system well enough to be able to speak to the viability of the suggested changes, however, I do note though that some countries with a centralised approvals process can also incur delays of up to 2 years so this proposed change per se may not be a central answer. The most important thing is that I think this paper, which describes the problems of usual modus operandi, needs to be contrasted with the current processes at the height of the pandemic. One of the silver linings of the pandemic is how quickly the system has changed and how many large (and often previously underused) datasets have

	been brought into service so that critical public health questions can be raised and answered in literally one or two weeks using data at a national level. The real question in this space is not how we fix the previously dysfunctional system but what will the post-pandemic health data landscape look like?

REVIEWER	Raman, Sudha R. Duke Univ
REVIEW RETURNED	19-Apr-2021

GENERAL COMMENTS	In this Communication format manuscript, the authors detail the path to obtaining the necessary approvals and governance to linking 5 datasets about people with CHD. The authors then make some recommendations about possible steps that researchers and other stakeholders could undertake to improve the processes, with the ultimate goal to pave the way for a more efficient way to use the wealth of information held in these data sources. The specific examples were very informative – I think many researchers may not realize who many are experiencing similar challenges with data linkage. Several additions could improve the communication's utility. 1) A limitations section nearer to the end of the paper. Understanding that this is single study experience, it would be a good addition to tell the reader in what ways this study is similar to previously studies, and what factors may reduce the applicability of this experience. 2) In the discussion/conclusion, the paper should include a strengths section, in particular recognizing that research communications such as this are important to disseminate information about the practice of research, and aligns with various recommendations about the feasibility of research using linked data. Placing the paper in some context that would be helpful to the reader – some examples may be https://pubmed.ncbi.nlm.nih.gov/28369581/, https://pubmed.ncbi.nlm.nih.gov/31950565/, there are likely more are region specific. 3) The language is very informal at times, which is engaging for the reader but perhaps not specific enough (soul destroying, top tips) 4) The tables and figures are helpful convey the sheer magnitude of the study activities. One additional figure to summarize the recommendations may be helpful. 5) Minor comments: Perhaps add more detail or a reference for the following statements: a. "Patient representatives on a different data linkage project that experienced similar challenges….. b. "Failure to comply with the data protection principles could mean a fine of up to €20 million, or 4% of an organisation's total worldwide annual turnover, whichever is higher c. Please clarify or explain this this underlined section: CAG have now introduced their precedent set pathway14 for an expedited review, which although a step in the right direction requires time-limited access to undertake record linkage/validation and anonymisation of data. Thank you for the opportunity to review this paper.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Dermot OReilly, Queens University Belfast
Comments to the Author:

Dear Editor

Manuscript ID bmjopen-2020-047575.R1, entitled "The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement

Thank you for asking me to review this interesting and well written paper. I especially liked the title, though they may also want to consider…'the long and winding road' or 'Chasing the rainbow's end'…

We thank Dr OReilly for these positive comments. We thought about changing the title but decided to stick with the original version as "the road to hell is paved with good intentions" is a common, well known phrase.

This is an unusual paper and the authors have done well to try and make it fit into the usual academic paper format.

I entirely understand and appreciate the frustration and exasperation that they must have felt for much of this journey, and I think the picture will be well recognised by many in the field. Indeed, there are some of us who have been trying to access particular datasets from unnamed Departments for almost seven years.

I have no real problems with the paper and think it should be published both as a potential road map for others to follow and a reassurance to others that their case is not unique; though it may deter some from embarking on the journey at all! I would ask the authors whether they thought their decision to ask individual data controllers to link their data to NCHDA identifiers was a food one and may have added to their woes and troubles?

Thank you for these supportive comments. Originally we were planning on asking NHS digital to perform the linkage for all datasets. We have added the following section under 'Ethics & CAG', as part of our explanation of the CAG precedent set pathway and how this wasn't applicable to us.

> 'The original data linkage plan had involved each audit sending their data to NHS Digital, so we would receive the final linked dataset only. However, following feedback from a study that had adopted this strategy, that this had lengthened the process, without them having access to any of the datasets during this time, we opted for each audit completing it. This allowed us to receive and begin working on each dataset as it reached us, without needing to have received every dataset. NHS Digital was the lengthiest application process and therefore the original planned linkage would have delayed us even further. We did not seek to receive the identifiers from each audit for the sake of data minimisation and governance.'

There are many important points here such as the need for understanding funders and the difficulties of identifying in advance how much time the various processes will take. I do not know the English system well enough to be able to speak to the viability of the suggested changes, however, I do note though that some countries with a centralised approvals process can also incur delays of up to 2 years so this proposed change per se may not be a central answer.

We thank for the reviewer for this comment and have added detail around this point.

> 'Another possibility being adopted in other countries[18 19] is a central authority for processing and approving applications, and implementing linkage, thereby reducing the number of data controllers carrying out effectively the same review, and increasing efficiency. However, some countries adopting such systems have reported that such systems are not without delays and problems and that further improvements are required. [19 20 21]'

The most important thing is that I think this paper, which describes the problems of usual modus operandi, needs to be contrasted with the current processes at the height of the pandemic. One of the silver linings of the pandemic is how quickly the system has changed and how many large (and often previously underused) datasets have been brought into service so that critical public health questions can be raised and answered in literally one or two weeks using data at a national level. The real question in this space is not how we fix the previously dysfunctional system but what will the post-pandemic health data landscape look like?

We thank for the reviewer for this suggestion and have added a sentence reflecting on the pandemic experience.

> 'Lessons from the pandemic
>
> During 2020, research governance was relaxed in pursuit of rapid scientific evidence into COVID-19 aetiology, risk factors, and treatments. This included a fast track review for ethics review,[22] the pause of the need for approval under Regulation 3(4) of the Health Service Control of Patient Information Regulations 2002,[23] and the release of data and change in process by some data controllers.[24] This expedited progress of all studies into COVID, and prioritisation of COVID research, shows that there is room to simplify the process.[10] This provides an important opportunity to learn from what happened during the COVID pandemic, and what could be adapted for the future.'

Best wishes
Dermot

Reviewer: 2
Dr. Sudha R. Raman, Duke Univ
Comments to the Author:
In this Communication format manuscript, the authors detail the path to obtaining the necessary approvals and governance to linking 5 datasets about people with CHD. The authors then make some recommendations about possible steps that researchers and other stakeholders could undertake to improve the processes, with the ultimate goal to pave the way for a more efficient way to use the wealth of information held in these data sources. The specific examples were very informative – I think many researchers may not realize who many are experiencing similar challenges with data linkage.

We thank Dr Raman for these positive comments.

Several additions could improve the communication's utility.
1) A limitations section nearer to the end of the paper. Understanding that this is single study experience, it would be a good addition to tell the reader in what ways this study is similar to previously studies, and what factors may reduce the applicability of this experience.

We agree this would be a useful addition and have added a limitations section in the discussion.

> **Strengths & Limitations**
>
> Our experience of an ambitious project to link five datasets may not be entirely generalisable to other researchers, especially those applying for fully anonymised data from an established research database.
>
> However for most studies, although some differences would apply for ethics and CAG requirements, such requirements will still need to be considered at the outset of the study. Datasets required will differ for different projects, but we expect that each research team will encounter similar governance requirements albeit with different requirements of the relevant data controllers. Longitudinal studies and clinical trials may have many approvals in place, but

may be still experiencing their own challenges with changing governance requirements that could delay renewal of permissions.

2) In the discussion/conclusion, the paper should include a strengths section, in particular recognizing that research communications such as this are important to disseminate information about the practice of research, and aligns with various recommendations about the feasibility of research using linked data. Placing the paper in some context that would be helpful to the reader – some examples
may be https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubmed.ncbi.nlm.nih.gov%2F28369581%2F&data=04%7C01%7C%7C4f9c46f7403c4a253a7208d91395eb7e%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C637562356912799587%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=CnfwJtNR9Q41eE8bg1kA5hlkYxTUrwwsJ9tM2666oMo%3D&reserved=0, https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fpubmed.ncbi.nlm.nih.gov%2F31950565%2F&data=04%7C01%7C%7C4f9c46f7403c4a253a7208d91395eb7e%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C637562356912799587%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=PS5LA74LIjcRVk3RKGk7inwVi6glUjgGI2WOWFTxQrM%3D&reserved=0, there are likely more are region specific.

We agree this would be a useful addition and have added a strengths section in the discussion with the limitations (as above), followed by the following paragraph to place the paper in context.

> 'Our challenges were not unique,[7 11 13 25] and align with other recommendations that have been suggested.[8 10 13 26 27] The publication of our experience is an important addition to the research method dissemination, hopefully facilitating the development of new and more efficient data access systems.'

3) The language is very informal at times, which is engaging for the reader but perhaps not specific enough (soul destroying, top tips) [Editor's note: although the language doesn't need to be overly formal, please do consider whether non-native English speakers will understand terms such as 'soul destroying'. We have a very large international readership.]

We thank the reviewer and editor for these comments. We have changed the term 'soul destroying' to 'demoralising and seemingly hopeless', and the term 'top tips' to 'useful advice'.

4) The tables and figures are helpful convey the sheer magnitude of the study activities. One additional figure to summarize the recommendations may be helpful.

We thank the reviewer for this suggestion, and agree this would be useful to readers. We have added a figure 3 to illustrate our suggested improvements for the future.

5) Minor comments: Perhaps add more detail or a reference for the following statements:
a. "Patient representatives on a different data linkage project that experienced similar challenges…..
b. "Failure to comply with the data protection principles could mean a fine of up to €20 million, or 4% of an organisation's total worldwide annual turnover, whichever is higher
c. Please clarify or explain this this underlined section: CAG have now introduced their precedent set pathway14 for an expedited review, which although a step in the right direction requires time-limited access to undertake record linkage/validation and anonymisation of data.

We thank the reviewer for these comments. For a, the comments were from project team meetings so we have deleted this sentence, we have added a reference for b and a more detailed explanation for c. Please find this explanation below:

‘CAG have now introduced what is known as the precedent set pathway[15] to enable a more timely review process, which although a step in the right direction applies only to specific situations, known as the -‘precedent set categories’. These categories include: applications to identify a cohort of patients and subsequently seek their consent; accessing data on site to extract anonymised data; validity of consent; data cleansing of historical studies; time limited access to undertake record linkage/validation and anonymisation of data,

The latter category was not a feasible category for us to apply for expedited review as we required pseudonymised data from each data controller to link all datasets at UCL. The original data linkage plan had involved each audit sending their data to NHS Digital, so we would receive the final linked dataset only. However, following feedback from a study that had adopted this strategy, that this had lengthened the process, without them having access to any of the datasets during this time, we opted for each audit completing it. This allowed us to receive and begin working on each dataset as it reached us, without needing to have received every dataset. NHS Digital was the lengthiest application process and therefore the original planned linkage would have delayed us even further. We did not seek to receive the identifiers from each audit for the sake of data minimisation and governance.’