International Journal of Population Data Science

Journal Website: www.ijpds.org

Identifying researcher learning needs to develop online training for UK researchers working with administrative data: CENTRIC training

Fiona Lugg-Widger[1*], Kim Munnery[1], Julia Townson[1], Rob Trubey[1], and Michael Robling[1,2]

Submission History

Submitted:	22/10/2021
Accepted:	13/12/2021
Published:	02/02/2022

[1]Centre for Trials Research, Cardiff University, Cardiff, CF14 4YS
[2]DECIPHer - Centre for Development, Evaluation, Complexity and Implementation in Public Health Improvement, 1-3 Museum Place, Cardiff. CF10 3BD

Abstract

Background
The use of administrative data in health and social science research continues to expand, with increased availability of data and interest from funders. Researchers, however, continue to experience delays in access, storage and sharing of administrative data. Training opportunities are limited and typically specific to individual data providers or focussed on the analytical aspects of working with administrative data. The CENTRIC study was funded by the Information Commissioners Office, with the aim of developing a broader training curriculum for researchers working with administrative data in the UK.

Methods
A mixed-methods design informed curriculum content, including surveys with researchers, focus group discussions with data providers and workshops with members of the public. Researchers were identified from relevant administrative data networks and invited to participate in an online survey identifying training needs. Data providers were approached with a request to input to a face-to-face or online meeting with two members of the research team about their experiences of working with researchers. Data were analysed within the broad framework of the interview schedule, free text responses in the survey were analysed thematically.

Results
107 researchers responded to the online survey and four data providers participated in the focus groups. We identified five main themes, relating to research training needs for UK researchers working with administrative data: communication; timelines; changes & amendments; future-proofing applications; and, the availability of training and support. Data providers either provided additional evidence on these learning needs or ways to address identified challenges. Six modules were developed addressing these training needs. Quotes from the survey and focus groups are used anonymously in the online training modules.

Conclusion
The CENTRIC online training curriculum was launched in September 2020 and is available, free of charge for UK researchers. CENTRIC specifically addresses commonly identified training needs of researchers working with administrative data.

Keywords
administrative data; routine data; training; data access

*Corresponding Author:
Email Address: LuggFV@cardiff.ac.uk (Fiona Lugg-Widger)

https://doi.org/10.23889/ijpds.v7i1.1712

Introduction

The use of administrative data for health and social care research continues to expand [1–3]. In health research, the UK Health Data Research (HDRUK) Alliance is leading the way in bringing together data, expertise and infrastructures enabling health research in the UK [4]. NHS DigiTrials is one of the initiatives born out of HDRUK, which aims to streamline trials applying for data through NHS Digital in England and offering support to clinical trials such as recruiting potential participants [5]. The use of administrative data for social science research is also expanding. In England, administrative data research (ADRUK) and Data First have facilitated linkages between the Department for Education and Ministry of Justice, enabling criminal justice, education and social care data to be linked for the first time [6]. Funding initiatives encouraging the use of administrative data are also increasing [7].

Despite the increased availability of data and interest from funding bodies, there remain challenges for data access, storage and sharing, and difficulties in conducting cross-national research [2, 8–11]. In addition, accessing data from multiple data providers remains a challenge with different applications, requirements and governance across providers [9]. Processes and legal frameworks also change over time, as is expected, with post-GDPR changes having most recently moved the goal posts for researchers applying for data [12].

The UK Clinical Research Collaboration (UKCRC), National Institute for Health Research (NIHR) and Health Research Authority (HRA) have historically attempted to bridge the gap between researchers and data providers, facilitating discussions and change informed by those experiencing the challenges [13–15] However, applications to access and process data shared by UK data providers are usually undertaken by researchers working alone or in small teams. This is evidenced for example, by the large number of data security protection toolkits listed for users of NHS Digital, often within the same university [16]. This fragmentation means that despite some central coordination by university research governance departments, much onus remains on individual teams in gaining access to and processing data. Problems accessing, storing and sharing data continue to be experienced in silos [11]. Data providers work with a large number of data recipients. Indeed, the breadth of research topics and projects that request administrative data means each application presents its own nuanced challenges making guidance and support challenging to address by each data provider.

To date, there have been limited training opportunities for health and social care researchers to learn about the challenges and opportunities involved in preparing, applying for and working with administrative data. Some data providers require researchers to undertake some form of training prior to working with administrative data. Mandated training invariably relates to safe researcher training (data protection, information security, confidentiality) and use of data safe havens specific to one data provider. Other established training (delivered by academic organisations) for researchers is heavily focused upon technical challenges such as data cleaning, standardisation and models of data linkage (i.e., probabilistic, deterministic linkage techniques) [17, 18].

The Information Commissioner's Office (ICO) funded the CENTRIC project in 2019 to address this gap. The aim was to provide a broader training course that would better support researchers in understanding and navigating the process of gaining access to routine public sector data, and to help them process it in a regulatory compliant manner [19]. Importantly, the content of the UK-focused CENTRIC training course was to be informed through consultation with researchers, data providers and members of the public. Aligning with the ICO strategy, this funded work addressed three of their strategic goals: (1) increasing public trust and confidence in how data are used and made available; (2) improving standards of information rights practice through clear, inspiring and targeted engagement and influence; and (3) staying relevant, providing excellent public service and keeping abreast of evolving technology [20].

This paper sets out the methods applied to identify the key areas of training needs as identified by researchers and data providers. Methods of co-production with members of the public are to be published separately.

Methods

This was a mixed-methods study designed to understand learners' needs from the perspective of both researchers and data providers, in order to co-produce a training package to address specific learning outcomes linked to the ICO strategic objectives [20]. Ethical approval for the study was provided by Cardiff University School of Medicine Research Ethics committee (Ref 19/51). The terms *routine data* and *administrative data* were used interchangeably during this study, with administrative data being used in the final training curriculum.

Researcher survey

We developed a survey aimed at UK researchers who work with administrative data to identify training needs and training preferences. The aim was to use the learning from these surveys, along with feedback from data providers, to develop an online curriculum with two face-to-face workshops.

The survey comprised four sections: (1) respondent characteristics; (2) a training needs assessment; (3) information about personal learning style; and (4) researcher experience of public involvement. These domains were selected to enable the team to understand who was completing the survey (i.e., their level of experience in routine data), the key areas a curriculum should cover to identify the desirability of different learning styles to ultimately be included for both the online course and face-to-face workshops and to explore current experiences of public involvement in routine data research aligning with the ICO strategic goals to increase trust and confidence.

The survey was piloted in June 2019, with eight researchers, five of whom participated in a de-brief interview. These individuals were accessed via a local routine data network of academics. The pilot ensured the flow of questions was acceptable, including checking the logic (i.e., skips) and user feedback helped to identify questions that could be worded better and topics felt to be missing. The time taken

to complete the pilot surveys were recorded, to inform future users of an expected completion time.

The main survey was hosted by Online Surveys (https://www.onlinesurveys.ac.uk) and respondents provided their consent before starting the survey. We used a non-probability convenience sampling approach with the intention to include researchers from a range of disciplinary backgrounds and across different public sectors but not with the aim of producing generalisable findings. We expected that we may attract in excess of n=100 respondents. Recruitment for the survey was via social media, e-mail distribution lists (e.g. AllStats, UK Trial Manager Network and UK CRC) and via local contacts in this field. The survey was open for 10 weeks. Survey questions were made up of both open and closed questions. Quantitative data were analysed descriptively. Participant responses in the free text boxes were analysed thematically. Codes were recorded using NVivo (version 12) and themes identified by authors (FLW, KM). All data for the survey are stored within Online Surveys secure data centres operated by Amazon Web Services and Cardiff University servers.

Focus Groups with UK data providers

We held focus groups with key UK data providers between July – September 2019, both face-to-face and by videoconference. We approached key staff at five data providers, who were involved in data access requests, from England, Wales and Scotland. We targeted those data providers who were considered national administrative data providers for researchers covering the health and social care sectors, informed by the literature [21]. Staff within each data provider were invited to take part in these focus groups, provided with an information sheet and those interested then contacted the study team and a mutual date and time was identified. Two researchers with experience in conducting focus groups (FLW with KM or MR) facilitated the sessions, which each followed the same topic guide. Topic guides were focused on the context of developing a training curriculum for UK researchers, and covered: a description of the data providing organisation and their data access process; challenges and successes of data request applications; available training; the role of the public; monitoring and evaluation of their data provision (audits and at the service level). These areas aligned to both the outputs of the researcher survey and the ICO strategic goals. Consent was recorded via consent forms and confirmed verbally at the start of each session. All focus groups were recorded and transcribed verbatim. Transcripts were uploaded into NVivo (version 12).

Analysis (Survey and focus group)

We analysed survey free text responses and the focus group transcripts using framework analysis [22] based on the structure of the questions. This was performed by KM and FLW respectively. Other themes that emerged outside of the framework were identified and coded as such.

Curriculum development

Data from the survey and the focus groups were reviewed and discussed by the study team. The data gathered were used to both inform the scope and content of the training but also used as evidence in the training (i.e., quotes) to emphasise or explain a point.

Results

Survey respondent characteristics

A total of 107 responses were received. Not all questions were mandatory, so denominators are provided for each question. 88/99 (89%) respondents were based in a university and the remaining 11/99 (11%) were based at an NHS or Government agency. A breakdown by country of employer is available in Table 1. 95% (82/87) had worked with health administrative data but other data were represented also (Table 2).

Data provider characteristics

Twelve individuals working within data access roles from four data providers contributed to the focus groups. These data providers represented health, education and social care data from England, Scotland and Wales (NHS Digital, Department for Education, electronic Data Research and Innovation Service and the Secure Anonymised Information Linkage Databank).

Themes

There were five high level themes that appeared in both the focus groups with data providers (DP) and the survey with researchers (RS). These are illustrated by a sample of quotes in text and in Supplementary Appendix 1. Quantitative survey data supporting these themes are also presented.

Theme 1: communication

The first and most prominent theme from both the survey and focus groups was communication challenges. Two issues were apparent through the survey and focus group: a difference in conceptual understanding (i.e., understanding of some of the requirements involved in data applications); and terminological differences (the same term being consistently used differently or being understood differently). Both researchers and data providers mentioned speaking different languages and at times, at crossed purposes. This challenge, perceived by both parties, resulted in delays on both sides while meetings and subsequent drafts of data request applications were (re-)reviewed to ensure mutual understanding.

> *The different language [Data Provider] uses and learning it. Understanding what they mean in the questions on their application form. They do provide support, but it would be good to get a very good first draft prepared before needing the help. (RS44)*

> *"We are essentially people who are from different backgrounds looking at things from different perspectives, and trying to communicate in ways that we can both understand." (DP1)*

Table 1: Country of employer[1]

	N (%)
England	65 (60.1%)
Northern Ireland	<5 (<5.0%)
Scotland	13 (12.1%)
Wales	16 (14.9%)
UK wide[2]	<5 (<5.0%)
Missing	9 (8.4%)

[1]Respondents provided *the name of current main employing organisation* which has been aggregated to country.
[2]Organisations which worked across more than one nation.

Table 2: Data types respondents reported accessing

Type of data	N (%)[2]
Health (incl. registries)	82 (95.3)
Office for National Statistics (ONS)	34 (39.5)
Education	13 (15.1)
Social Care	11 (12.8)
Criminal Justice	8 (9.3)
Third / Charitable Sector	5 (5.8)
Work and Pensions	2 (2.3)
Other[1]	4 (4.7)

[1]Other included transport, housing, environment and census data (Scotland).
[2]Multiple responses were recorded per respondent.

We explored with data providers how best to address this challenge. From their perspective, the key challenge was researchers being able to simplify their language for non-expert readership, to ensure a shared understanding of the project and methods.

Theme 2: timelines

Although linked to communication, timelines appeared as a separate theme for both researchers and data providers including expectations of the process and related timelines for data access requests. Exploring this from both sides of the application process, it is clear that delays are incurred, and a source of frustration, for both parties.

> *"It is also challenging when you hear nothing / no updates for a period of time and have no information as to whether the application is progressing as it should or it's been forgotten about (we have experienced both)." (RS96)*

> *"An application is submitted, and we try and speak to the researcher to refine the form, and discuss the governance panel, and for one reason or another, they become quite incommunicado. So the forms can't be progressed if we can't have that engagement with the researcher... And not having that two-way engagement all the time slows down the process...it's not always as easy to... hold their place in the queue, because while we're waiting for them we're processing other cases ... because the delay, the impact on your own research project, it's an impact on other projects as well that we're trying to, trying to process." (DP3)*

Researchers regularly referred to a lack of transparency from data providers on such aspects as timelines and the likely time required to approve applications. From a data provider perspective, they had to manage what they felt were often unrealistic expectations on the speed of approval.

We explored with data providers how best to identify realistic timelines, and how they felt that researchers could potentially mitigate delays in data request applications that impacted time-sensitive aspects of their research. All four data providers emphasised the importance of researchers engaging with their application team as early as possible in the project. They all felt that this would address expectations on timescales, and identify any obvious stop-go criteria that would need to be addressed before proceeding with data request approval.

Theme 3 – changes and amendments

Another challenge related to timelines were the changing requirements of a research project over the course of the project. In part, this reflects the changing nature of research as new information / research comes to light, but does also highlight the need for the research team and data providers to engage at an earlier stage to agree datasets and associated costs.

> *"we do come across projects where they started off and they want three data sets, and by the*

time they come to permissions they're looking for an additional six or seven data sets linked into that study. Um that can cause some interesting conversations if they've gone for funding based on the original." (DP4)

Data providers felt that significant changes could often have been avoided by discussing the data flows and datasets earlier in the study process. Designing data flows to maximise regulatory compliance requires input from all parties (data providers, regulators and the sponsor). A data flow set-up for one project may not be appropriate for all future study designs, and as technology continues to advance, different methods of data transfer and storage may become a preferred option for data providers.

"As a department we don't necessarily want to feel backed in or forced into a route. We want to work with, what's the most secure way and most efficient way of achieving your project's um desired outcome." (DP3)

This quote also raises co-production of data sharing approaches and the importance of early discussions with those involved in the release of data.

Theme 4 – future-proofing the application

A significant challenge, and an area in which the needs of researchers and data providers could often diverge, related to the length of time that the research data could be held, for archiving and/or re-use in further future research. Consent considerations for both of these scenarios, as well as associated costs require up front consideration. Consent wording will need to be approved at the data request application, and costs agreed at grant application stage. Researchers felt that this is not always feasible or realistic.

The survey asked researchers "What key challenges do you face in **planning** a study using routine data?":

Lack of clarity about what is required on consent forms to link trial data to routine data in future (especially non-health data) (RS89)

How long you are allowed to keep data, in order to plan whether several applications for extension will be needed over the course of a project or not. (RS5)

Costs over time change (e.g. costs for archiving is now a consideration whereas before it did not need to be costed) (RS107)

This highlights the ongoing challenge for researchers to put things in place at the start of a project that will still be relevant years down the line. Time between study set-up and linkage may well be one aspect that can be managed in future projects by conducting more than one linkage as the below data provider describes:

"I would always say to research projects: don't leave it so long . . . people that tend to do this better will not collect stuff for 5 years and then try and link it. They'll collect for 6 months and they'll do a trial linkage so we're covering consent here but also just also in terms of data quality and linkage variables knowing what your collecting, is collected in the right way and the right format, is it linkable, have you got errors all these little things. . . it will help you out in the long term, rather than failing really hard at the end of the process." (DP1)

One challenge (for both researchers and data providers) is when regulations change rendering earlier decisions incompatible with current regulations. The impact of General Data Protection Regulation (GDPR) on historic trials (i.e. set-up and consented years before the DPA changes were in development) remains a challenge today. Data providers and researchers continue to navigate this hurdle. Considerations related to section 251 support to enable a legal flow of data is a useful option for those working with health-related data, and awareness of this as an option seems important to enable research projects to continue their work.

Theme 5 - training for researchers

When asked about formal routine data training they had completed, one third of researchers (26/77) indicated they had completed no formal training related to working with routine data. 43% (n = 33) stated training such as the Medical Research Council (MRC) Research, GDPR and confidentiality training [23] and the ONS Safe Researcher Training [24] – which are mandatory for those accessing data from some data providers. Data provider specific training (including workshops and webinars) accounted for 13% of the training listed, in-house training was mentioned by 17% of researchers and 13% stated having received training on the statistical methods applied to routine data. Informal support researchers received primarily came from colleagues (43%), Data Providers (16%) and networking/conferences (19%). Multiple answers were possible for both of these questions.

The focus groups with data providers explored training that they make available, or recommend to those applying for data. Much of their training was in the form of guidance documents available on their website or sent to applicants alongside the application form or phone calls with applicants. Two data providers noted challenges relating to their guidance not being read when applications were being submitted and one data provider highlighted the problem of researchers often relying on informal training, or seeking support from colleagues who had previously applied for data under different governance arrangements.

"So in my experience I feel as though a lot of the experience in academic research is handed down. So we get new researchers coming through, PhD students etc, and they speak to people who've done research before, which obviously makes sense. And they might be told, 'we use this consent form' . . . and then they'll use that consent form virtually word for word with just the particulars of their project passed on. And actually they're not getting the current guidance. . . Somebody who's done research for 40 years will say what you do is you get the data,

Box 1: Data provider training wishlist

- Process of applying (flow diagram)
- Interdependencies on getting an application approved (ethics, security, etc)
- Opportunity for applicants to see data before requesting (dummy data)
- Required preparation before applying
- How to write a good application (using the right language)
- Safe haven (what it is, how it can be used)
- Disclosure control
- How to have a successful two-way conversation with data providers
- GDPR concepts / ICO requirements
- Understanding the legislative framework
- Privacy notices
- Timelines / project management
- Data management (recording what you're doing, decisions made)

> *and you store it here and so on. Whereas actually I think it's worth people considering whether they actually need to have the data there, whether that's the optimum way forward, whether it's the most cost-effective way forward. But also as part of that decision they should be talking to the data providers, because there'll be an awful lot that people don't know. People won't necessarily know about [accessing data via a safe haven] for example. And their mentors in the profession won't necessarily know that either, despite our best efforts to try and get the information out there." (DP2)*

Data providers reported providing support to the research community via local research networks, roadshows, conference workshops and workshops hosted at the organisation. Future plans included webinars, animations and videos to be made available on their website. Data providers were specifically asked what they would want to see in a training curriculum to address the challenges they experience with applicants, and also made suggestions throughout the focus groups. These were thematically coded and Box 1 presents all suggestions identified from the discussions.

Discussion

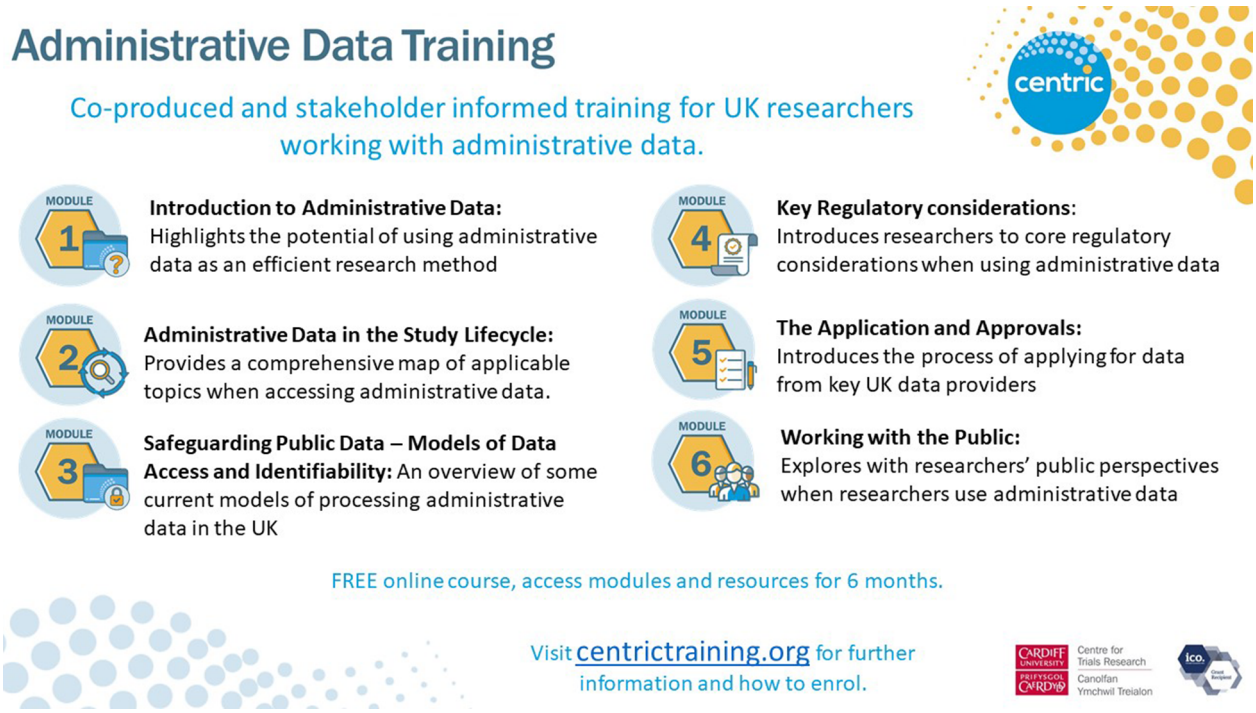
Summary of findings

Five main themes identified areas for improvement from which we elicited specific learning needs for UK researchers working with administrative data. These were: Communication; Timelines; Changes & amendments; Future-proofing applications; and, (lack of) available training and support. Although presented as two separate themes due to the wealth of data, communication and timelines both contribute to the problem of the other (i.e. poor communication leads to increased timelines and unrealistic expectations of timelines is due to poor communication). In a similar vein, changes to the study design or datasets can impact timelines of the study approvals and could be avoided with better communication, earlier in the study. The experiences and challenges of researchers attempting to future proof their applications are echoed by the data providers. Indeed, changing regulations have impacted both parties and the required changes to processes and data application requests. The availability of training for researchers for working with administrative data primarily focusses on the secure access and processing of such data or technical methodological approaches to analysing these data. Informal support from colleagues was the most relied upon form of training, however, data providers highlighted the risk in this approach (i.e. providing out-dated information) and the potential negative impact on data requests.

Other ongoing work

Many challenges highlighted here have been anecdotally reported in relation to specific projects or data providers [9, 11, 25]. Through this work, we have identified the broader current issues relating to working with administrative data, as well as developing a more detailed understanding and context around known issues by working with both researchers and data providers. Funding has been directed to address some of these ongoing limitations of administrative data with a report being issued by Department of Health and Social Care to review the current state of the use of health data for research and analysis (Goldacre Review[26]). The Health Data Research UK (HDR UK) was established in 2018 as the national institute for data science in health. Through the UK HDR Alliance (launched in February 2019) and funding

Figure 1: CENTRIC training module details



from UK Research and Innovation and medical charities such as the Wellcome Trust and the British Heart Foundation, HDR UK Hubs and HDR Gateway, HDR UK brings together health data assets with specialist expertise across academia, industry and healthcare [4, 27]. HDR UK have recently published, via HDR UK Futures, a range of bitesize videos across a range of curriculums including analytical skills, using the HDR UK Gateway, health data access and PPIE for health data research [28]. The Trials Methodology Research Partnership (TMRP), funded by NIHR and MRC, are also progressing the methodological challenges through the Health Informatics working group and routine data topic group [29]. Both national networks are addressing broad health-related and trial-related challenges respectively. At a more local level, data providers are reviewing their application forms, processes and communication with researchers.

Strengths and limitations

Aligned with the ICO strategic goals [20] we have developed a public-informed training programme through a process of co-production which explores public views and public understanding, tailored to the needs of researchers. We sampled researchers from many disciplines and levels of experience in using administrative data as well as four of the largest data providers in the UK. To note, this did not include a data provider in Northern Ireland (NI) and unique features of the data access arrangements and customers for NI data may not be well represented in our current curriculum. We will seek to address such potential gaps in future iterations of the training. Through co-production with members of the public we have ensured the "working with the public" module in particular is relevant and addresses areas of importance felt by the public. This training was developed pre-pandemic, and experiences of data access reflect a pre-pandemic view. Access to data has changed, in particular for COVID-19 related projects receiving expedited approvals causing delays for other (non-COVID-19 related) research. Processes in place to deliver COVID-19 research rely on the suspension of Regulation 3(4) of the Health Service Control of Patient Information Regulations 2002 which is only a temporary situation during pandemic times [30]. Calls for these changes to be implemented to the post-pandemic business-as-usual procedures have been supported by over 350 researchers in the UK [31, 32]. This may impact some elements of the CENTRIC curriculum and highlights a key challenge in this ever-evolving area of research – the need for regular review to ensure sustainability and ongoing relevance.

The CENTRIC training curriculum

Six modules were agreed upon to cover the content of training (Figure 1). Across the six modules developed for this training curriculum, all of the five mentioned themes and challenges are addressed and included. The training includes tips, checklists and examples throughout the curriculum to improve communication between researchers and data providers, highlighting common pitfalls and quotes from the focus groups. Modules 1 and 2 provide an overview of how accessing administrative data fits into the study lifecycle and notes key points in which delays can occur and ways to avoid this. To reduce the need for amendments, Modules 3 and 4 provide considerations to data storage and applicable regulatory considerations that may result in changes to aspects of study design (e.g., data flows). Module 5 provides detailed information on the application process which aims to familiarise the learner with this process and set realistic expectations related to timelines. Other relevant training courses are linked to throughout the course, where these fall

outside the remit of the curriculum: for example, statistical and data management skills. Public input and engagement is covered in Module 6, reflecting the importance as well as the practical challenges of public input and engagement in administrative data research. This final module was co-produced by members of the public and the development will be reported separately. The online course went live in September 2020. Face-to-face training days were planned for the summer of 2020, however, this was moved to online webinars (*Getting a grip! Regulatory requirements when using administrative data; Involving the public in studies using administrative data – the How and the Why.*) held in October 2020 due to the COVID-19 pandemic.

To assess impact, in the short-term, we evaluated the training curriculum in two ways. Firstly, within the online training course, learners were invited to provide module specific feedback via a brief survey. Secondly, we encouraged reflection and feedback from participants who took part in our online webinars, to assess whether the content of the online course (and the webinars themselves) matched the needs of the learners. For the longer-term, we are engaging with data providers, regulators and funders for feedback and endorsement. We will also continue to monitor the feedback being provided by learners in the course modules, to ensure that the content continues to meet researcher and data provider needs.

Conclusion

The CENTRIC online training curriculum has been developed using a co-production model informed by the public, researchers and data providers and is available, free of charge, for UK researchers. The training addresses the most commonly identified challenges and training needs, as described in this paper, for both researchers and data providers. We hope that this training course will facilitate improved access and management of administrative data in the UK and help reduce barriers that we have identified between researchers and data providers.

Acknowledgments

We would like to thank all those who took part in the surveys, workshops and focus groups as part of this study. The Centre for Trials Research receives funding from Health and Care Research Wales and Cancer Research UK. The study was funded by the Information Commissioner's Office.

Statement on conflicts of interest

The authors declare that they have no competing interests.

Ethics Statement

Ethical approval for the study was provided by Cardiff University School of Medicine Research Ethics committee (Ref 19/51)

References

1. Hemingway, H., Lyons, R., Li, Q., Buchan, I., Ainsworth, J., Pell, J. and Morris, A. (2020) "A national initiative in data science for health: an evaluation of the UK Farr Institute", *International Journal of Population Data Science*, 5(1). https://doi.org/10.23889/ijpds.v5i1.1128

2. Lee, S., Xu, Y., D'Souza, A. G., Martin, E. A., Doktorchik, C., Zhang, Z. and Quan, H. (2020) "Unlocking the Potential of Electronic Health Records for Health Research", *International Journal of Population Data Science*, 5(1). https://doi.org/10.23889/ijpds.v5i1.1123

3. Mc Cord KA, Al-Shahi Salman R, Treweek S, Gardner H, Strech D, Whiteley W, et al. Routinely collected data for randomized trials: promises, barriers, and implications. Trials. 2018;19(1):29. https://doi.org/10.1186/s13063-017-2394-5

4. UK Health Data Research Alliance. Available at https://ukhealthdata.org/. Accessed 22 Oct 2021.

5. NHS Digitrials. Available at https://digital.nhs.uk/services/nhs-digitrials. Accessed 22 Oct 2021.

6. ADRUK website: https://www.adruk.org/our-work/browse-all-projects/data-first-harnessing-the-potential-of-linked-administrative-data-for-the-justice-system-169/

7. UKRI ESRC: https://www.adruk.org/news-publications/funding-opportunities/

8. Jones KH, Heys SM, Daniels H, Ford DV. Exploring barriers and solutions in advancing cross-centre population data science. Int J Popul Data Sci. 2019 Aug 5;4(1):1109. https://doi.org/10.23889/ijpds.v4i1.1109. PMID: 34095536; PMCID: PMC8142621.

9. Lugg-Widger FV, Angel L, Cannings-John R, Hood K Hughes K, Moody G and Robling R (2018) Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass. *IJPDS Special issue: Cross-Centre Working*, 3:3:2, https://doi.org/10.23889/ijpds.v3i3.432

10. National Cancer Research Institute. The researchers' experience when attempting to access health data for research. Available at: https://www.ncri.org.uk/accessing-health-data-for-research/ Accessed 22 Oct 2021.

11. Macnair, A., Love, S.B., Murray, M.L. *et al.* Accessing routinely collected health data to improve clinical trials: recent experience of access. *Trials* **22**, 340 (2021). https://doi.org/10.1186/s13063-021-05295-5

12. Russell, A.E., Ford, T., McIntosh, A., Jones, P.B., Shenow, S., Russell, G. and McManus, S., 2021. Researcher access to mental health data: results from an online consultation.

13. Cross, L., Carson, L. E., Jewell, A., Heslin, M., Osborn, D., Downs, J. and Stewart, R. (2020) "Guidance for researchers wanting to link NHS data using non-consent approaches: a thematic analysis of feedback from the Health Research Authority Confidentiality Advisory Group: A thematic analysis of feedback from the Health Research Authority Confidentiality Advisory Group", *International Journal of Population Data Science*, 5(1). https://doi.org/10.23889/ijpds.v5i1.1355

14. NHS Digital Event Review: 25 November 2020 Using health and social care datasets in research. Available at: https://digital.nhs.uk/services/research-advisory-group/using-health-and-social-care-datasets-in-research Accessed 22 Oct 2021.

15. Applebe, D., Parker, C.,and Hartley, S. Making Requests to NHS Digital Easier. Available at: https://www.nihr.ac.uk/documents/explore-nihr/Efficient%20studies/Liverpool%202017%20Final%20Report-Making%20Access%20to%20NHS%20Digital%20Easier.pdf Accessed 22 Oct 2021

16. Data Security and Protection Toolkit. Organisation Search. Available at: https://www.dsptoolkit.nhs.uk/OrganisationSearch Accessed 22 Oct 2021.

17. University of York. Analysing Patient-Level Data using Hospital Episode Statistics (HES). Available at: https://www.york.ac.uk/che/courses/patient-data/ Accessed 22 Oct 2021.

18. University College London. Introduction to Hospital Episode Statistics. Available at: https://www.ucl.ac.uk/child-health/events/2021/oct/introduction-hospital-episode-statistics Accessed 22 Oct 2021.

19. Information Commissioner's Office. Cardiff University. Available at: https://ico.org.uk/about-the-ico/what-we-do/grants-programme/cardiff-university/ Accessed 22 Oct 2021.

20. Information Commissioner's Office. Our mission, vision, strategic goals and values. Available at: https://ico.org.uk/about-the-ico/our-information/mission-and-vision/ Accessed 22 Oct 2021.

21. Lensen, S., Macnair, A., Love, S.B. et al. Access to routinely collected health data for clinical trials – review of successful data requests to UK registries. Trials 21, 398 (2020). https://doi.org/10.1186/s13063-020-04329-8

22. Gale, N.K., Heath, G., Cameron, E. *et al.* Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Med Res Methodol* **13**, 117 (2013). https://doi.org/10.1186/1471-2288-13-117

23. MRC. e-Learning. Available at: https://byglearning.com/mrcrsc-lms/course/index.php?categoryid=1 Accessed 29 Nov 2021.

24. Office for National Statistics. Accessing secure research data as an accredited researcher. Available at: https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme Accessed 29 Nov 2021.

25. Taylor JA, Crowe S, Espuny Pujol F, *et al* The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement*BMJ Open* 2021;**11**:e047575. https://doi.org/10.1136/bmjopen-2020-047575

26. GOV.UK. New review into use of health data for research and analysis. Available at: https://www.gov.uk/government/news/new-review-into-use-of-health-data-for-research-and-analysis Accessed 22 Oct 2021.

27. HDRUK. Our funders. Available at: https://www.hdruk.ac.uk/about-us/funders/ Accessed 29 Nov 2021.

28. Health Data Research UK. Continued Professional Development. Available at: https://www.hdruk.ac.uk/careers-in-health-data-science/continued-professional-development/power-up-your-health-data-science-knowledge/ Accessed 22 Oct 2021

29. Trials Methodology Research Partnership. TMRP Working Group Funding Awards https://www.methodologyhubs.mrc.ac.uk/about/tmrp-working-group-funding-awards/

30. Department of Health and Social Care. Coronavirus (COVID-19): notice under regulation 3(4) of the Health Service (Control of Patient Information) Regulations 2002. Available at: https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information/coronavirus-covid-19-notice-under-regulation-34-of-the-health-service-control-of-patient-information-regulations-2002-general–2 Accessed 29 Nov 2021.

31. Cavallaro, F., Robling, M., Lugg-Widger, F., Cannings-John, R., Aldridge, R., Gilbert, R., Harron, K. Open letter Open letter to the ICO, CMOs and UK data providers: Reducing barriers to data access for research in the public interest-lessons from covid-19 (with signatories). Available at: https://eprints.ncl.ac.uk/file_store/production/267907/3FF575A5-4795-40EA-B0EB-447290C11B3E.pdf Accessed 22 Oct 2021.

32. Cavallaro, F., Lugg-Widger, F., Cannings-John, R., Harron, K. 2020. Open letter to the ICO, CMOs and UK data providers: Reducing barriers to data access for research in the public interest-lessons from covid-19. *British Medical Journal*.

Abbreviations

ADRUK : administrative data research
CENTRIC : Co-Produced and stakeholder informed training for UK researchers working with routinely collected data
GDPR : General Data Protection Regulation
HDRUK : Health Data Research UK
HRA : Health Research Authority
ICO : Information Commissioner's Office
MRC : Medical Research Council
NIHR : National Institute for Health Research
TMRP : Trials Methodology Research Partnership (TMRP)

Supplementary Appendix 1: Supplementary quotes from researchers and data providers

Theme	Supplementary quotes
Communication	"As an example, everybody in the research environment seems to understand what it means to sponsor a study, and I've had numerous definitions of this and still don't have a concrete view." (DP2)
	"It's the ability to step outside what you know, and view the same subject from the perspective of a lay person. . . it's literally how you explain what you do to the man in the street . . . we see a lot of applications with an awful lot of Latin in, or very detailed information about the statistical analysis. . ." (DP2)
Timelines	"Time. Time. Time. It is almost impossible to even roughly estimate a timeframe for a project using routinely collected data. Information Governance approvals, data sharing agreements, and data transfer are all areas in which the researcher has no direct control over timeframe." (RS58)
	"And that's one of our challenges that we face, where we'd receive an application form in, and [a researcher says] 'I need my data in two weeks' time'. You say well sorry but it takes two weeks to create and curate the data, never mind to get it through the governance process." (DP1)
	So it's about engaging with us as a data sharing team at the earliest opportunity. You know, there's some fantastic research ideas out there, and some fantastic people who want to do this research. But pressures of actually getting the data in a timely manner, you know, and talking to us first, and saying "I want to do this research, from your experience do you think it would be something that [data provider] could help with? How long might it take to get that data?" But also about, not assuming you can automatically have the data. It's not a given that you can have our data (DP3)
Changes and amendments	"It could be, we see this commonly with section 251, people go and get section 251 support, and then we're actually not happy with the methodology, because it involves us flowing more data than is actually needed to. Or sometimes receiving a huge amount of data from a third party, such as an auditor or a register to link, whereas actually all we need is the identifying data, we don't need the clinical data, and then we can provide a bridge file. So if we can liaise on things like methodology at the earliest point, we can address a lot of those issues." (DP2)
Future-proofing the application	"I think consent as a concept changes over time and some of it can be handled on best intent at the time of collection of consent. But there are hard and fast rules that people can't get around now with things like DPA, GDPR things like that so it's just having that knowledge and being aware that when you are starting to collect data or when you want to start using it for research purposes, do you do things with best intent? Do you consult with the relevant data providers or whoever they are to get it as close as possible to correct?" (DP1)
	"A lot of the studies are follow-up studies, so whilst they may have been having access to health records during that period, after a certain amount of years they're then going to come into us to follow them up, to have the data. And it's then that the consent model is a lot out of date, and it just doesn't meet today's standards or requirements. And that's a big issue. . . It's making that decision as well as whether, if there's enough for it to stand as to meet the common law of duty, or whether you would then have to go and get section 251. And I think for some of our researchers that can be seen as. . . another obstruction, but it's not, it has to meet, you know, it's a legal requirement that you do have to have a legal basis to be able to have access to that". (DP2)
Training for researchers	"The aim with the [guidance documents] is to firstly improve transparency with our customer base, so that they can understand this is what this part of the application form has to adhere to in order to meet the approval standard." (DP2)
	"we offer a service where if people want a telephone call to discuss their application, or discuss the requirements of their data share, we're quite happy to speak to people. Or if people want a hand filling in the application form, we'll also talk through application processes" (DP3)
	"what we've done is we've written a guidance document, made that available within our environment for people to refer to. But you can lead a horse to water, but you can't always get it to drink the water." (DP4)
	"the amount of organisations that don't read the guidance that we put out." (DP3)