International Journal of Population Data Science

Journal Website: www.ijpds.org

Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass.

Lugg-Widger, FV[1], Angel, L[1], Cannings-John, R[1], Hood, K[1], Hughes, K[2], Moody, G[1], and Robling, M[1]

Submission History

Submitted:	24/11/2017
Accepted:	22/05/2018
Published:	21/09/2018

[1]Centre for Trials Research, Cardiff University
[2]Division of Population Medicine, Cardiff University

Abstract

Introduction
Researchers are increasingly using routinely collected data in addition to, or instead of, other data collection methods. The UK government continues to invest in research centres to encourage use of these data, and trials and cohort studies utilise data linkage methods in the follow-up of participants. This does not come without its limitations and challenges, such as data access delays.

Objective
This paper outlines the challenges faced by three projects utilising individual-level routinely-collected linked data for the longer-term follow-up of participants.

Methods
These studies are varied in design, study population and data providers. One researcher was common to the three studies and collated relevant study correspondence, formal documentary evidence such as data sharing agreements and, where relevant, meeting records to review. Key themes were identified and reviewed by other members of the research teams. Mitigating strategies were identified and discussed with a data provider representative and a broader group of researchers to finalise the recommendations presented.

Results
The challenges discussed are grouped into five themes: Data application process; Project timelines; Dependencies and considerations related to consent; Information Governance; Contractual. In presenting our results descriptively we summarise each case study, identify the main cross-cutting themes and consider the potential for mitigation of challenges.

Conclusions
We make recommendations that identify responsibilities for both researchers and data providers for mitigating and managing data access challenges. A continued conversation within the research community and with data providers is needed to continue to enable researchers to access and utilise the wealth of routinely-collected data available. The suggestions made in this paper will help researchers be better prepared to deal with the challenges of applying for data from multiple data providers.

Focus of this article

- Routinely-collected data are increasingly accessed in health and social care research however there are challenges for researchers working with multiple data providers
- This paper outlines the challenges faced by three studies accessing data from a number of data providers in England and Wales
- Recommendations, based on the experience of these studies, are made for researchers to consider and incorporate into future studies accessing routinely collected data

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

http://dx.doi.org/10.23889/ijpds.v3i3.432

Introduction

The use of routinely-collected data in health research is increasing and expanding to enable more efficient trial designs at a reduced cost (1–4). Using data available on populations on a national, European and global scale creates an opportunity to benefit health research in academia and beyond (5). Over recent years, the UK government and devolved administrations have invested in a number of e-health and administrative data networks to encourage innovations in this area, including the £100m investment of the Medical Research Council (MRC) since 2012 into awards for initiatives such as the Farr Institute. Similarly, the MRC and other UK funders recently announced their commitment of £37.5m over five years to "*transform the UK medical informatics research landscape*" through Health Data Research UK (6).

In a trial setting, typical risks and potential biases to a project include participant recruitment and retention. Routinely-collected data can identify potential study subjects and reduce data collection required from participant, thereby addressing some of these. However, it may also add risk, for example, in gaining appropriate access. Accessing routinely-collected data has previously been described as "virtually impenetrable terrain, hostile to the research pioneer" (5). Delays in accessing data from providers can be due to the legal landscape governing the use of data for research and the diverse frameworks implemented by data providers to support decision-making for data applications (5,7). These challenges are well documented (8) and despite calls for improvement (9) researchers still struggle to access data for studies (10). The legal and ethical aspects of accessing, using or linking health data have received more critical attention than other administrative data, such as social care (11–13). Beyond the frustrations and time taken to access data, these challenges have wider impacts on research. Lack of anticipated access to follow-up data reduces sample size and increase risk of bias, loss of confidence by funders in organisations to deliver research, and wasted money spent on accessing data rather than analysis, interpretation and knowledge dissemination.

Using our recent experience from three studies, across different populations and data providers (Health, Education and Social Care), we aim to summarise some key challenges, lessons learned and solutions in accessing routinely collected data for the longer-term follow-up of trial and cohort participants.

Methods

We used a narrative case study approach to identify themes and suggest solutions to arising challenges. Three studies were selected opportunistically as examples of current work undertaken by the same research centre and involving use of data from multiple data providers. We included studies that are varied in design, study population and data providers (Table 1). For each we summarised in tabular form key features of the study design, particularly aspects related to routinely-collected data. One researcher was common to the three studies (RCJ) and FLW led on collating relevant information. This included discussions with each of the respective research teams (including the Chief Investigators: MR, KHu), review of relevant study correspondence (e.g. emails), formal documentary evidence such as data sharing agreements and, where relevant, meeting records. Each case study therefore represents the individual study design, application process and subsequent receipt of data.

The lead researcher identified key themes in a first stage of analysis. While not necessarily problems, these were challenges presented to the study teams involved. Summaries of challenges were reviewed by other members of the research team. In this process, potential and/or employed mitigating strategies were identified. To ensure that the perspective was not solely that of the research teams involved, the draft summary was shared with a data provider representative (GC). This provided an opportunity for validation from a key external stakeholder and importantly a means to establish the feasibility of suggested mitigation strategies (for example where there was interface with the data centre). Finally, a broader group of researchers (all from the same research centre) with an interest in research using routinely-collected data were invited to comment on a draft summary document.

First, we describe the three contributing studies:

The Building Blocks Trial

The Building Blocks Trial (BB:0-2) was a Randomised Controlled Trial (RCT) assessing the effectiveness of the Family Nurse Partnership (FNP) home-visiting programme when added to usual care, as compared with usual care alone, for young mothers and their first child living in England. The trial utilised prospective data collection along with use of routinely-collected data from birth records, GP practices, abortion data, death data from the Office for National Statistics (ONS) and hospital data from NHS Digital (NHSD) (at the time, Health and Social Care Information Centre) (14,15). Participants provided their explicit consent for data linkage.

The Building Blocks follow-on study

The Building Blocks: 2-6 Study (BB:2-6) involves the same participants as the BB:0-2 trial, but with a primary focus on maltreatment. The follow-up of this cohort is solely via routinely-collected data and is without explicit consent (opt-out model supported by Section 251 approval). Data include abortion data (16), ONS data, hospital data from NHSD (17), education data and social care data from the National Pupil Database (NPD)(18).

The LUCI Study

The LUCI Study (The Long-term follow-up of Urinary Tract Infection (UTI) in Childhood) is an electronic record-linked study following up two cohorts of children living in England and Wales who consulted a GP with acute illness when aged 5 or under during either the DUTY (19) or EURICA study (20). Follow-up data will include hospital and GP data from the Secure Anonymised Information Linkage (SAIL) Databank (21–23), Microbiology culture data from Public Health Wales (via SAIL) and hospital data from NHSD.

Data Providers

The data providers discussed in this paper are as follows:
NHS Digital (NHSD) is the Data Controller for clinical data on patients in England, for example from hospital records, cancer registrations, patient demographics and also mortality data available from the Office for National Statistics (ONS)(17). The Department for Education is the Data Controller for the NPD, which includes information sourced from English publicly-funded schools, local authorities and awarding bodies (24). SAIL Databank provides clinical data on patients in Wales including from GP, hospital and the Welsh demographic service. SAIL also act as a data safe haven, providing secure storage and access to data via their remote portal. The Department of Health is the Data Controller for the data held within the Abortion Statistics Team (AST) provide data on terminations of pregnancies to 'bona fide researchers' for approved requests (16). See Table 2 for a summary of the Data Providers' application and publishing requirements.

Other key committees / organisations

The Confidentiality Advisory Group (CAG) provides independent expert advice to the Health Research Authority on the use of confidential patient information (25). The UK is governed by the Information Commissioner's Office, an independent body set up to uphold information rights as per data protection legislation (26).

In presenting our results descriptively we will summarise each case study, the main cross-cutting themes identified and the potential for mitigating challenges identified.

Results

The identified challenges are collated under five headings: Data application process; Project timelines; Dependencies and considerations related to consent; Information Governance; and Contractual. Table 3 summarises the challenges and potential mitigating actions.

1. Data application process

The key challenges identified under this theme are changes of requirements over time, the length of the application process and the differing requirements among UK data providers.

Study approval process

Until recently, NHSD would not accept an application until the ethical approval letter, and other approvals such as legal approval for linking data (section 251), were uploaded as part of the application. For some projects, NHSD application approval relied on Section 251 support, which in turn relied on ethics approval; one application could not proceed until the previous requirement was fulfilled. The same was true for any substantial amendments required to the study. NHSD are now working with Health Research Authority and other organisations to identify areas for improvement which, includes allowing approvals to be submitted in parallel rather than sequentially (18).

Changes over time

We experienced a number of changes to application forms, approval processes and review panels (members and remit) during the lifecycle of the three studies. These changes in governance reflect data providers' ensuring data releases are in the public benefit and comply with data protection law and ethical standards. In the case of NHSD, changes were also a response to the PriceWaterhouseCooopers Data Release Review Audit overseen by Sir Nick Partridge (27,28) and to public concerns about care.data (29). This in turn extended timelines for the re-drafting and re-submission of a data request application and delays from the data providers whilst staff and panel adapted to the changes and communicated with the applicant accordingly.

Length of application process

In our three studies, the duration of time between submission and signing the data sharing agreement (at which point data can be transferred between the two organisations) ranged from two months to 18 months depending on the data provider and when the data were requested (e.g. during a period of change as above). Table 1 outlines the various data providers and associated timelines.

For BB: 2-6, once the application was in the new template NHSD progressed it to the review panel within a reasonable timeframe (three months) in February 2015 and three points were noted as the reason for not approving the application. Once these were addressed, instead of re-review by the panel, the internal review process then identified further queries/comments over a period of 11 months (two periods of change, six rounds of internal review, three different case officers assigned) preventing its recommendation to be considered by the NHSD Panel until February 2016.

The queries and comments raised by the internal (pre-panel) review were incorporated into the BB:2-6 application and where relevant informed the application for the LUCI study. A combination of these lessons learned (Box 1), no period of change, and the continuity of one case officer throughout the process enabled the LUCI study application to receive a four-and-a-half-month turnaround from submission of a completed application to signed data sharing agreement (enabling data transfer).

NHSD listened to criticisms raised by the researcher community and have responded with further guidance, webinars and increased communication. Communication via audio conferences between the research team and NHSD, in particular with the case officer presenting the application to the panel and the data analyst preparing the data, also enabled greater understanding of the application and a faster response to queries raised.

Differing requirements

The information required, including the level of detail, the application forms, and the review panel remit, differ across data providers. This is partly down to the nature of the data provider and the legal requirements placed on them regarding release of data. NPD for example are not required to comply with the Health and Social Care Act 2012 (30) whereas NHSD are, placing a focus on the applicant to demonstrate

Table 1: A summary of the three case studies – BB:0-2 Trial, BB:2-6 Study and LUCI Study.

	BB:0-2 Trial	BB:2-6 Study	LUCI Study
Ethical Approval Ref.	09/MRE09/08	14/WA/0062	16/WA/0166
Legal basis for linking data[1]	Explicit consent	s2513 Ref. CAG 10-08(b)/2014 s42(4) of the Statistics and Registration Service Act 2007 6 (1) of Schedule 2 of the 1998 Data Protection Act	s251 Ref. 16/CAG/0114
Primary Outcome(s)	Subsequent pregnancy; Smoking in late pregnancy; Emergency hospital attendance and admissions; Birthweight	Child in Need status	Renal Scarring
Funder	Department of Health Policy Research Programme.	National Institute for Health Research, Public Health Research Programme (NIHR PHR)	Health and Care Research Wales
Population	Young mothers and their first-born child(ren)		Children <5 years presenting with acute illness in Primary Care
Study design	RCT	RCT with longer term follow-up using routine data	Cohort with longer term follow-up using routine data
Location of data for analysis	Cardiff University server	SAIL data safe haven	SAIL data safe haven
Data providers & Timelines[2]			
NHS Digital (including ONS)	5 months (2013-2014)	18 months (2014-2016)	4.5 months (2017)
Abortion Statistics team	1 month (2012)	2 months (2017)	N/A
National Pupil Database	N/A	7 months (2016)	N/A
SAIL	N/A	N/A	5 months (2016)

[1]At the time of approval; [2]The duration of time between submission and signing the data sharing agreement (at which point data can be released); [3]Section 251 of the Health and Social Care Act 2012.

Table 2: Overview of the relevant Data Providers

	NHS Digital	ONS (via NHSD)	NPD	SAIL	Abortion Statistics Team
Approval of data application	Independent Group Advising on the Release of Data (IGARD Panel)		Tier 1 data[1]: Data Management Advisory Panel Tier 2-4 data[2]: Education Data Division	Information Governance Review Panel	Chief Medical Officer
Evidencing IG[3] requirements	Data Security and Protection (DSP) Toolkit (Previously IG Toolkit)		Information Security Questionnaire	Remote Access only	Study-specific basis
Data access cost	Available on Website		No Cost	Available on Request	No Cost
Approval timelines	3 Tier system (15; 30 & 60 days) stop-clock		Not available	Estimated 12 weeks	Not available
Contractual	Organisation Framework agreement				
Project level Data Sharing Agreement (DSA)	As per NHSD + Individual Declarations	Project level DSA Individual Declarations	Person level User Agreement	Project level DSA	
Publication requirements	Acknowledgement	Acknowledgement Disclosure control for cell counts <3	Acknowledgement Notification prior to publication	Acknowledgement Notification on acceptance of publication 3 references provided for inclusion Disclosure control for cell counts <5	Acknowledgement Notification prior to publication Disclosure control for cell counts <9
Maximumt duration of contract	3 years	1 year	3 years	Project-specific (and within the funded period for SAIL)	3 years

[1]Tier 1 data are the most sensitive personal information; [2]Tier 2, 3 & 4 data are other sensitive personal information, school level data and other pupil-level data (respectively); [3]Information Governance – the processing of data in line with required standards.

measurable benefit to the public that NPD are not required to consider.

Navigating the numerous data request applications and required documentation was a challenge for all three studies as it requires more than copying the information from one application into another. NHSD plans to update their requirements for those studies funded by National Institute for Health Research (NIHR), for example where a clear patient benefit is a requirement for funding, and thus has already been adjudicated. This will allow some applications to progress quicker but is not a solution for studies that are funded by other organisations.

The key mitigation activities during the data application process are: 1) for researchers to understand any process changes by monitoring data provider(s) website(s), signing up to receive email notifications/newsletters, and attending related webinars; 2) for data providers to manage information about their processes and to communicate any changes to researchers; 3) for data providers to assign a single case officer to enable consistent communication and project knowledge; 4) for researchers to allow the time and resources for the data application process; and 5) for researchers and data providers to share knowledge of the different application forms to assist others in understanding what data providers are looking for in the application thereby reducing the number of re-drafts required prior to consideration by panel.

2. Project timelines

The unpredictability of how long it takes for an application to move through the data providers' process is affected by the above challenges and is in itself a challenge for the project. Delays in applying for and accessing routinely-collected data from multiple data providers poses a significant risk to project delivery. There is also a wider risk that public funding is wasted on project extensions due to delays in data request applications. This risk may deter funders from supporting projects that rely heavily on data from external providers.

For BB:2-6, we included a feasibility stage within the project where we requested example data from data providers prior to the final extraction. This enabled the team to assess data quality and prepare cleaning and analysis scripts. However, in the context of this paper, it particularly informed the research team about their subsequent data applications (for example, wording and documentation) to enable a faster timeframe than previously experienced. Staff resource reflected the anticipated workload, with reduced resource assigned during periods of low activity and increased once data were available. We communicated with both the study funder and the independent study steering committee throughout this feasibility phase, to discuss the impact of application delays on study deliverability. Assessing the feasibility of accessing data from multiple data providers meant we could provide reassurance to the funders.

The LUCI study benefitted from the lessons learnt from BB:2-6, updated the wording of the participant material, created a participant facing website and included further detail in the data application request on processing, benefits and output. Other actions such as updating the University required data protection notification to the Information Commissioners Office (ICO) and ensuring review of the IG Toolkit were also in place prior to this application.

Mitigation for data application timelines include research teams drafting the applications, discussing with the data providers, and talking to other teams of researchers who have experience working with the data provider. More generally, assessing feasibility is an important element to consider and communicate. There is, for example, a NPD user group. Sharing of applications would also benefit other researchers; although project details will be different, the underlying principles will be the same. This would show the language used in the applications which would help other researchers understand how to describe the research in sufficient detail but at a level of understanding required by a review panel.

3. Dependencies and considerations related to consent

The key issues identified here related to fair processing and the amount of detail provided to participants. Fair processing refers to the requirement, as per UK Data Protection legislation, to be transparent about how the data collected will be used, stored and shared (25).

The data used for research are changing and so is the public's relationship to privacy (31). It is important to ensure participants, whose personal data we are processing for linkage purposes, are adequately informed about how their data will be used for research. There are a number of models of consent for health information (31) all of which fit under different legal bases for the transfer personal information used for data linkage (e.g. explicit consent, s251 support). Ensuring data providers agree with the wording of participant information sheets and consent form under their "fair processing" checks was a challenge for all the projects.

In the example of BB:0-2, NHSD did not initially accept the wording around the linkage element included in the participant information sheet and consent forms used for study recruitment. NHSD recommended modifications to the study consent, and that all the BB:0-2 participants be re-consented using the modified wording. Re-consenting would likely introduce non-response that would impair the validity of the sample, and in particular lead to the loss of some of the most vulnerable participants. Recruiting to the study was a considerably challenging task, as was maintaining adequate follow-up at the required assessment time points; these were, indeed, key rationales for obtaining routinely-collected data for the study. Whilst we agree that meaningfully informing participants about research access to their records is valid, there is a challenge in providing sufficient information to potential participants about data linkage when also presenting other required information (e.g. potentially submitting to an intensive two-year intervention). Following an appeal, the application was approved. The data provider maintained their preference for alternate wording but recognised the potential damage to the study re-consenting could cause.

NHSD will soon be introducing a nine-point check which the panel will use to assess fair processing notices (e.g. is the information published, is it clear and truthful, what data will be collected?). This will enable trials which hold data from applications made before new requirements came in place, such as BB:0-2, to continue to hold data, providing these nine considerations are adequately addressed.

Box 1 Lessons learned from addressing NHS Digital requirements (BB:2-6 study)

Demonstrating benefit to health and social care

- Referencing relevant government policies and priorities to evidence the importance of the research
- Describing how results will be disseminated to reach policy, practice and academic stakeholders to enable a benefit to health and social care.

Clarity of planned data flow

- Highlighting critical transitions and other elements in data flows, for example, where data move from identifiable to de-identified, who has access to the data, where data are held as individual level or aggregate,

Clarity about personnel involved in data processing

- Providing essential description of staff (and organisations) who access the data, and for what purpose (i.e. data loading, data cleaning, data analysis)

Assuring compliance with Data Protection legislation

- For example, amending organisational notification to ICO for Universities processing the data to explicitly state data processing for health/healthcare data and analysis/research (previously only top-level Research had been selected in notification).
- Update to Ethics and CAG approvals so that participant facing material explicitly stated the data providers, data requested, who will access the data and why.

Confirmation about funding source

- Letter from funder to confirm details not explicit in the confirmation of funding letter for example that ONS data will be required for the project and study start/end dates.

Assuring Information Governance

- Ensuring all stages of the data flow pathway are supported by appropriate evidence (for example, submission from SAIL to demonstrate their IG compliance; letter from Director of Research and Innovation Services, Cardiff University, to confirm contracted requirements placed upon SAIL, Swansea University are in place)
- Ensuring IG Toolkit (now DSP) is maintained as up-to-date and that self-assessment has been reviewed.

Clarity about legal basis

- Highlighting how contracts held with NPD demonstrate legal basis to link to NPD data.
- Providing supplementary letter from CAG to confirm details not already explicit in confirmation letter.

This challenge was addressed in the BB:2-6 and LUCI study by having a participant-facing study website. This was the main information mechanism for those who wished to opt-out but can also be used as a means of communication for long-term engagement of the cohorts and for updates on agreed changes to the processing of participant data.

During the development of study materials for all three studies the wording was reviewed by a lay group to ensure the information was communicated in a way that the public would understand. Ensuring a balance between providing participants with information that is sufficiently understandable but also comprehensive is a significant challenge. This may be particularly the case where routine data are used as this may be a relatively unfamiliar concept for members of the public. Ensuring all data providers are happy with wording also posed a challenge for BB:2-6. An agreed set of wording followed multiple reviews by the lay group, ethics panel, confidentiality advisory group (CAG), NHSD and NPD.

Mitigation strategies for issues related consent include flexibility of research teams in responding to questions during the application stage and updating documentation where required (including amendments to Ethics and CAG). Use of lay input can enable a strong justification for retaining the comprehendible participant-facing material and should continue to be the basis for content in participant facing materials. It is incumbent upon researchers to test the adequacy of draft participant materials to establish comprehension using approaches such as cognitive interviewing. Using other modes of communication (e.g. website) is also a good way to manage and required changes required later down the line and enhance longer term engagement.

4. Information Governance / Security

The key issue related to this theme is the challenge both to reach the required standard, and to provide the evidence of this. Information Governance (IG) can be defined as the process/es in place and controls that cover data collection, security of physical and electronic storage, use, data sharing, archiving, and ultimately destruction of data (32). IG is of upmost importance to all involved i.e. public, participants, data providers and research team. Evidencing IG becomes more complex when it involves data from multiple data providers (32) with multiple ways to show good practice, and multiple requirements placed on the data user.

NHSD require evidence primarily via the Data Security and Protection (DSP) Toolkit (previously called the IG Toolkit), an online system allowing a self-assessment against Department of Health standards to ensure processing of NHS patient data complies with UK legal frameworks (for example, Data Protection, common law duty of confidentiality (33), the Freedom of Information Act 2000 etc.)(34). NPD requires completion of a security questionnaire covering technical system details, physical security, data handling, and staff awareness. It does not recognise the DSP Toolkit, instead leaning more toward an organisation being ISO-27001 compliant (35). Some data providers such as the SAIL Databank do not allow data to be held outside of a specified secure environment. Security / IG evidence is not needed to access data from SAIL because SAIL holds the data and sets standards for all data users, such as protocols for remote data access. This set-up is useful in reassuring other data providers about data storage and access, but there are some drawbacks to this e.g. accessibility (printing and sharing of results in real time) and reliance on an external server to maintain disk space

While all of a similar merit to ensure data security, the different approaches pose a challenge to applicants requesting data. The resource implications are also an area for consideration. Completing and maintaining an DSP Toolkit registration is costly, especially when it is not recognised across all data providers, and ISO accreditation would be even more so. Interestingly, providing evidence that projects fall under a Clinical Trials Unit (CTU), are UK Clinical Research Collaboration (UKCRC) registered and all staff are GCP (Good Clinical Practice) trained, does not contribute to consideration of an application, and in some cases cause confusion when evidencing assurances around the processing of data.

Mitigation for information governance and security includes: 1) consideration of an organisation to be ISO 27001 certified or at least aligned has the benefit of a recognised IG standard; 2) appropriately resource time to implement the DSP Toolkit; and 3) ensure stakeholders in the broader organisation (e.g. outside of the applying department) are involved in planning / consultation. The use of data safe havens does mitigate some of the challenges, however the researchers will then need to choose between working with a limited subset of data in their own environment or potentially with richer data in restrictive settings that can hinder their productivity (21).

5. Contractual

Compliance of data access user agreements and retention of a licence to hold data over a number of years present challenges to university-level governance.

All data providers will issue a contract between themselves and the individual, project team and/or organisation the project team work within. The challenge with working with multiple data providers is the difference between the requirements placed on the organisation and the project team by these data providers. This includes IG but also covers data sharing, destruction, publication and retention.

Requirements for notification of publication, acknowledgement of data providers and inclusion of specific references in those publications differ across data providers, as do data retention periods and policies around cell suppression for small numbers (Table 2). Although many of these requirements are of similar nature, navigating these and ensuring all are adhered to does introduce a challenge to the project and indeed a risk, as these are contractual requirements.

For university research projects, there is a requirement to archive data for audit purposes. For example, Cardiff University requires research data to be archived for a minimum of 15 years. This puts a requirement on the research team and/or organisation to maintain active agreements with every data provider over 15 years to ensure archiving policies are upheld and that different data provider periods of retention are honoured (see Table 2). For some Data Providers a cost is incurred following each renewal of the contract (required to retain data) and this cost model has changed over the years. Renewal costs will not have been included in the study budget for projects conceived prior to these updated cost models, and research staff will change over 15 years. These issues are be-

ing discussed with NHSD for BB:0-2, with plans to pilot the extension of the data sharing agreement for archived data. Under a Data Archive Agreement, all NHSD data will be locked, restricting any further access (for example for the purposes of analyses or data checking). However, the agreement would cover a longer period and would be at an overall reduced cost compared to a data sharing agreement. If access to the data is required at any point, then an application to NHSD to allow processing (beyond archiving) will be required.

Ensuring compliance across data providers and documenting and evidencing these requirements was a challenge for all three studies. We have employed a member of staff who is responsible for ensuring policies and procedures in the department have consideration for these contractual requirements and that relevant studies document required evidence. In particular, documents such as the protocol, data management plan, statistical analysis plan and publication policy contain particular sections to address these requirements. These were developed in discussion with other universities and data providers. An additional mitigating strategy is to seek input from data providers to ensure studies are appropriately costed to enable long-term retention of data.

Discussion

Summary of the challenges

The challenges presented in this paper cover five themes. Changes to application processes and governance over time are inevitable and understandable. As law and public attitudes change, the availability of data and how researchers interact with them will need to adapt (32). There are calls for proportionate and principled governance (32) however this is likely to be data provider specific, ultimately resulting in differing principles for each data provider, as already experienced. Maintaining communication between applicants and data providers is key to mitigating delays where possible. The impact of delays on receiving data not only risks project delivery but also the development of staff. The rush to analyse and quality check data in the remaining funded time can limit the extent to which new staff can be trained in the use of these datasets, as well as limiting time to conduct exploratory data analyses to inform new projects and methods – a concern highlighted by others previously (7). Ensuring participants are informed about the study method and design in an understandable way is an important challenge. This is especially so when data linkage represents a small part of a larger complex RCT for which a participant will be providing consent.

At the Administrative Data Research Network (ADRN) conference in 2017 Julia Lane discussed the challenge of information being comprehensive (i.e. describes all necessary details for an individual to be fully informed about the data linkage process and privacy controls) and at the same time comprehendible (i.e. described at a level that a lay public member will read, understand and be able to use to make a decision). The conclusion was that it cannot be both (36) which reflects the challenges described here. Nevertheless, a greater exploration about what the public already understand and place importance upon in relation to routine data may aid researchers when making decisions about how to inform potential participants.

There remains more work to be done to enable researchers' access to data for appropriate research projects. The UK government published their review of data security, consent and opt-outs and highlighted the need to strike the right balance between privacy, confidentiality and data sharing (11); while privacy is very important, so too is data sharing (37). Currently, the re-use of such data for research requires a set of complex approvals from multiple governing entities which are often opaque, difficult to navigate and obtain, and so pose risks to population based research (9).

Strengths and Limitations

These three case studies represent the experience of three study teams within the same research centre. Although these experiences may not be reflective of all research studies accessing routinely collected data, they do offer learnings that will be applicable across trials and observational studies intending to access individual level routinely-collected data. Furthermore, two authors worked across all three studies providing a consistency in reviewing the challenges experienced. Nevertheless, this paper only focuses on the data providers applicable to these three studies. These are major providers of routine health and educational (including social care) data in the UK and will have relevance to a large number of researchers and studies. However, there may be other challenges and opportunities to learn from the experience of working with other data providers and so our own conclusions may be incomplete.

We did not specifically consider the challenges of analysing and interpreting routinely-collected data. These are important and valid concerns, which require on-going scrutiny but this paper focuses on the organisational and logistic aspects of multiple-provider working. Finally, we recognise that our reflections place greater emphasis on the perspective of the researcher rather than other agents, in particular the data providers. To address this, we shared our draft manuscript with a key contact within one of the provider organisations. We also worked closely with the providers in managing the process of applying for data, sometimes in the context of significant organisational change. Nevertheless, we accept that other issues may also be relevant which are not yet identified in our own report. This seems inevitable given the wide range of data users supplied by each provider and which may additionally underline the size of the task they too face in ensuring high standards of data governance.

Recommendations

When the Partridge review was published, approximately 30% of applications submitted to NHSD were applications from universities (27,28). Some of the challenges presented here may partially reflect that data users have very different interests and intended uses for the data and understanding the constraints researchers work within will be a useful message for data providers. Our key recommendations therefore relate to optimising communication at three different intersections within the research / data application context.

First, the research community and data providers need to continue their recent dialogue about data applications. The research community is not the primary user of routine health

Table 3: Challenges faced by the research team and proposed resolutions

Theme	Challenge	Resolution/Addressed how	Risk[1]
1. Data application process	Adapting to changes over time	Be aware of any changes by signing up to newsletters, email distribution lists.	Project
	Length of application process	Start discussions with data providers early on, factor in timelines at the funding application stage.	Project
	Different application requirements for difference centres	Resource this period of time appropriately and learn from other researchers.	Project
2. Project timelines	Unpredictability	Start discussions with data providers early on to be aware of additional delays.	Project
3. Dependencies and considerations related to consent	Ensuring Fair processing over long term	Consider using other methods of communication such as websites and seek data provider input early on to discuss acceptable options.	Project
	Comprehendible vs. comprehensive	Ensure documentation receives review from a lay representative and seek data provider input early on. Emphasis should be on it being understood. Formally test adequacy where possible.	Project
4. Information Governance	Differences across providers	ISO 27001 certification; Consider use of data safe havens to securely hold the data for the project. Appropriately resource time to implement the DSP Toolkit. Ensure stakeholders in the broader organisation (e.g. outside of the applying department) are involved in planning / consultation.	Project & Organisation
5. Contractual	Compliance with numerous Data Providers' contractual requirements	Development of standard operating procedures intrinsic to the department to address requirement.	Project & Organisation
	Long-term data retention	Appropriately cost studies to enable long-term retention. Seek input from data providers.	Project & Organisation
	Oversight of all requirements from all providers	Employ/fund a member of staff to take responsibility for ensuring policies and procedures in the department have consideration for all contractual requirements.	Organisation

[1] This refers to the level at which risk is implicated. Some challenges risk project delivery whereas others risk organisational compliance and reputation.

data but represent an important opportunity to add value to existing data to promote public benefit. Recent initiatives such as the UKCRC Registered Clinical Trials Unit workshop meeting with NHS Digital will help data providers better understand how data may be used and researchers better understand the constraints on data provision that may apply to their own requests.

Second, researchers using routine data need to share constructively their own experience of the data application process. User groups exist already (for example, for NPD data) although their focus may be more on analysis of the supplied data. This paper represents one example of helping others learning from our experiences and we hope others will be keen to do similar.

Third, funders and data providers themselves would benefit from greater dialogue. UK national funders such as NIHR may currently support a number of studies which are accessing routine data and will already have had fed back to them researchers' experience of the application process. As mentioned above, different regulatory requirements may apply to data providers such as NHSD and NPD, but some attempts to link data across NHS and Education show the interest in government departments in gaining research insights and value for money by sharing data. While harmonisation across data providers seems unfeasible, investing in attempts to identify and enhance commonalities of approach would likely pay important dividends for the UK.

Future directions / Areas of improvement

We highlighted some challenges that fall outside of the research team to mitigate. We therefore recommend that data providers provide and maintain communication around the application process and timelines especially when changes or delays occur. Our next recommendation relates to the challenge of archiving data and renewal of contracts, which is a risk for organisations. Data providers need to understand the wider research environment that researchers work within to understand how their contractual requirements conflict or agree with Good Clinical Practice (GCP) and Medicines and Healthcare products Regulatory Agency (MHRA) guidelines. Thirdly, we recommend that principles for data governance should be common across providers and greater movement should be made towards a single or exchangeable system. For example, if an organisation meets governance standards as determined by one data provider, this should provide a high level of reassurance across providers for working with routinely-collected data. The growing initiative for open and accessible data to be made available post publication is another challenge on the horizon for studies using routinely-collected data (38). We recommend data providers consider options to facilitate this.

A key unresolved issue from working with multiple data providers is how best to mediate conflicting requirements. To illustrate this, we used the example of participant information sheets and consent forms, which provide the basis for legal processing. Drafted participant materials may reflect the perspectives of researchers and guidance from lay representatives. However, they still may conflict with either ethics committees or requirements of data providers. In our experience funders and ethics committees place value upon materials that have meaningful lay input in their drafting and review. There are established methods to determine the quality of information provided to study participants and to improve their comprehensibility. We recommend that researchers should make greater use of these methods in preparing their study information. Data providers may have differing requirements for such information but these may not adequately account for the variation in experience, understanding and literacy that may be found in many study populations. Where study participants may be largely drawn from harder to reach populations this is especially important. We think that ultimately the public should be pivotal in establishing what information is provided and how. The researcher should facilitate this and provide evidence that such information meets standards for relevant comprehension.

Conclusions

Whilst some challenges eased in the years that these three studies were conducted, there remain a number of important areas for improvement. It is the responsibility of the research community to continue to identify these areas, within and outside of our control, to discuss and share experiences and solutions with other researchers with an objective of further improvement in processes for access to routinely-collected data. The recommendations made in this paper will help researchers better prepare for applications to multiple data providers and highlight potential modifications for data providers as well.

Acknowledgments

The Centre for Trials Research is funded by Health and Care Research Wales and by Cancer Research UK. We acknowledge and are grateful to Garry Coleman (NHS Digital Head of Data Access) for his input and support in this paper. We thank Elinor Coulman, Liz Merrifield, Jo Smith and Sarah Bridges for providing input as stakeholders in the development of this paper.

Funding

BB:0-2 trial was funded by the Policy Research Programme in the Department of Health (reference 006/0060). BB:2-6 study is funded by the National Institute for Health Research Public Health Research (NIHR PHR) Programme (reference 11/3002/11). The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR PHR Programme or the Department of Health. LUCI Study is funded by the Welsh Government through Health and Care Research Wales (reference 1068).

Statement on conflicts of interest

None declared

References

1. Harron K, Gamble C, Gilbert R. E-health data to support and enhance randomised controlled trials in the United Kingdom. Clin Trials [Internet]. 2015 [cited

2017 Sep 8];12(2):180–2. Available from: https://doi.org/10.1177/1740774514562030

2. Lewsey JD, Leyland AH, Murray GD, Boddy FA. Using routine data to complement and enhance the results of randomised controlled trials. Health Technology Assessment. 2000.

3. Facey K, Henshall C, Sampietro-Colom L, Thomas S. Improving the Effectiveness and Efficiency of Evidence Production for Health Technology Assessment. Int J Technol Assess Health Care [Internet]. 2015 [cited 2017 Sep 8];31(4):201–6. Available from: https://doi.org/10.1017/S0266462315000355

4. Cook JA, Collins GS. The rise of big clinical databases. British Journal of Surgery. 2015. https://doi.org/10.1002/bjs.9723

5. Laurie G, Sethi N. Towards Principles-Based Approaches to Governance of Health-related Research using Personal Data. Eur J risk Regul EJRR [Internet]. 2013;4(1):43–57. Available from: https://doi.org/10.1017/S1867299X00002786

6. Medical Research Council. Health Data Research UK (HDR UK) - About us - Medical Research Council [Internet]. 2017 [cited 2017 Sep 18]. Available from: https://www.mrc.ac.uk/about/institutes-units-centres/uk-institute-for-health-and-biomedical-informatics-research

7. Connelly R, Playford CJ, Gayle V, Dibben C. The role of administrative data in the big data revolution in social science research. Soc Sci Res. 2016; https://doi.org/10.1016/j.ssresearch.2016.04.015

8. Raftery J, Roderick P, Stevens A. Potential use of routine databases in health technology assessment. Health Technol Assess (Rockv). 2005;9(20). https://doi.org/10.3310/hta9200

9. Souhami R. Governance of research that uses identifiable personal data will improve if the public and researchers collaborate to raise standards. BMJ [Internet]. 2006 [cited 2017 Sep 8];333:315–6. Available from: https://doi.org/10.1136/bmj.333.7563.315

10. Powell GA, Bonnett LJ, Tudur-Smith C, Hughes DA, Williamson PR, Marson AG. Using routinely recorded data in the UK to assess outcomes in a randomised controlled trial: The Trials of Access. Trials [Internet]. 2017 [cited 2017 Aug 25];18:389. Available from: https://doi.org/10.1186/s13063-017-2135-9

11. National Data Guardian for Health and Care, National Data Guardian for Health. Review of Data Security, Consent and Opt-Outs National Data Guardian [Internet]. 2016 [cited 2017 Jul 11]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/535024/data-security-review.PDF

12. Sujan M, Howard-Franks H, Swann G, Soanes K, Pope C, Crouch R, et al. Impact of advanced autonomous non-medical practitioners in emergency care: protocol for a scoping study. BMJ Open [Internet]. 2017;7(1):e014612. Available from: https://doi.org/10.1136/bmjopen-2016-014612

13. Laurie G, Stevens L. The Administrative Data Research Centre Scotland: A scoping report on the legal & ethical issues arising from access & linkage of administrative data. Univ Edinburgh Sch Law Res Pap Ser [Internet]. 2014;35. Available from: https://doi.org/10.2139/ssrn.2487971

14. Owen-Jones E, Bekkers M-J, Butler CC, Cannings-John R, Channon S, Hood K, et al. The effectiveness and cost-effectiveness of the Family Nurse Partnership home visiting programme for first time teenage mothers in England: a protocol for the Building Blocks randomised controlled trial. BMC Pediatr. 2013;13:1. https://doi.org/10.1186/1471-2431-13-114

15. Robling M, Bekkers MJ, Bell K, Butler CC, Cannings-John R, Channon S, et al. Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): A pragmatic randomised controlled trial. Lancet [Internet]. 2016;387:146–55. Available from: https://doi.org/10.1016/S0140-6736(15)00392-X

16. GOV.UK. Abortion statistics, England and Wales [Internet]. [cited 2017 Oct 31]. Available from: https://www.gov.uk/government/collections/abortion-statistics-for-england-and-wales

17. NHS Digital. Data Access Request Service (DARS) [Internet]. [cited 2017 Oct 31]. Available from: http://content.digital.nhs.uk/DARS

18. Lugg-Widger F, Cannings-John R, Channon Sue, Fitzsimmons D, Hood K, Jones K, et al. Assessing the medium-term impact of a home-visiting programme on child maltreatment in England: protocol for a routine data linkage study. BMJ Open [Internet]. 2017 [cited 2017 Jul 17];7(e015728). Available from: https://doi.org/10.1136/bmjopen-2016-015728

19. Hay AD, Birnie K, Busby J, Delaney B, Downing H, Dudley J, et al. The Diagnosis of Urinary Tract infection in Young children (DUTY): a diagnostic prospective observational study to derive and validate a clinical algorithm for the diagnosis of urinary tract infection in children presenting to primary care with an acute illness. Health Technol Assess (Rockv). 2016;20(51). https://doi.org/10.3310/hta20510

20. O'Brien K, Edwards A, Hood K, Butler Christopher. Prevalence of urinary tract infection in acutely unwell children in general practice. BJGP [Internet]. 2013 [cited 2017 Aug 15]; Available from: https://doi.org/10.3399/bjgp13X663127

21. Lyons RA, Jones KH, John G, Brooks CJ, Verplancke J-P, Ford D V, et al. The SAIL databank: linking multiple

health and social care datasets. BMC Med Inform Decis Mak [Internet]. 2009 [cited 2017 May 5];9(9). Available from: https://doi.org/10.1186/1472-6947-9-3

22. Ford D V, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e- health research and evaluation. BMC Heal Serv Res BMC Heal Serv Res BMC Heal Serv Res [Internet]. 2009 [cited 2017 May 12];9(9). Available from: https://doi.org/10.1186/1472-6963-9-157

23. Jones KH, Ford D V., Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform [Internet]. 2014 [cited 2017 May 12];50:196–204. Available from: https://doi.org/10.1016/j.jbi.2014.01.003

24. GOV.UK. National pupil database: apply for a data extract [Internet]. [cited 2017 Oct 31]. Available from: https://www.gov.uk/guidance/national-pupil-database-apply-for-a-data-extract

25. Information Commissioners Office. No Title [Internet]. [cited 2018 Apr 4]. Available from: https://ico.org.uk/about-the-ico/who-we-are/

26. Authority HR. No Title [Internet]. [cited 2018 Apr 4]. Available from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/

27. Partridge N. Review of data releases by the NHS Information Centre [Internet]. 2014 [cited 2017 Oct 31]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/367788/Sir_Nick_Partridge_s_summary_of_the_review.pdf

28. PricewaterhouseCoopers LLP. Data Release Review [Internet]. 2014 [cited 2017 Oct 31]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/367791/HSCIC_Data_Release_Review_PwC_Final_Report.pdf

29. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. J Med Ethics [Internet]. 2015 May 23 [cited 2017 Oct 31];41(5):404–9. Available from: https://doi.org/10.1136/medethics-2014-102374

30. Health and Social Care Act 2012 [Internet]. Queen's Printer of Acts of Parliament; 2012 [cited 2017 Oct 31]. Available from: http://www.legislation.gov.uk/ukpga/2012/7/contents/enacted

31. Upshur R, Morin B, Goel V. The privacy paradox: laying Orwell's ghost to rest. Can Med Assoc. 2001;165(3):307–9.

32. McGrail K, Buchan I, Plankey Nathan Christopher Lea M, Nicholls J, Dobbs C, Sethi N, et al. Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. JMIR Med Inf. 2016;4(2). https://http://doi.org/10.2196/medinform.5571

33. NHS Digital. [cited 2018 Apr 4]. Available from: https://digital.nhs.uk/codes-of-practice-handling-information

34. Department of Health. Information Governance Toolkit [Internet]. [cited 2017 Oct 31]. Available from: https://www.igt.hscic.gov.uk/

35. Department for Education. National pupil database: apply for a data extract - GOV.UK [Internet]. 2014 [cited 2017 Nov 14]. Available from: https://www.gov.uk/guidance/national-pupil-database-apply-for-a-data-extract

36. Lane J. A call to action to build research data infrastructure. Nat Hum Behav [Internet]. 2017 [cited 2017 Oct 31];1:75. Available from: https://www.nature.com/articles/s41562-017-0075.pdf

37. National Advisory Group on Health Information Technology in England. Making IT Work: Harnessing the Power of Health Information Technology to Improve Care in England [Internet]. 2016 [cited 2017 Jun 19]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/550866/Wachter_Review_Accessible.pdf

38. Open Data | The BMJ [Internet]. [cited 2017 Nov 14]. Available from: http://www.bmj.com/open-data

Abbreviations

ADRN	Administrative Data Research Network
AST	The Abortion Statistics Team
BB:0-2	The Building Blocks Trial [ISRCTN23019866]
BB:2-6	Building Blocks: 2-6 follow-on study
CAG	Confidentiality Advisory Group
CTU	Clinical Trials Unit
DSP	Data Security and Protection Toolkit (previously IG Toolkit)
FNP	Family Nurse Partnership
GCP	Good Clinical Practice
IAO	Information Asset Owner
ICO	Information Commissioners Office
IG	Information Governance
LUCI	The Long-term follow-up of Urinary Tract Infection (UTI) in Childhood Study
MHRA	Medicines and Healthcare products Regulatory Agency
MRC	Medical Research Council
NIHR	National Institute for Health Research
NHSD	NHS Digital - The Health and Social Care Information Centre
NPD	National Pupil Database
ONS	Office for National Statistics
RCT	Randomised Controlled Trial
SAIL	Secure Anonymised Information Linkage
UKCRC	The UK Clinical Research Collaboration
UTI	Urinary Tract Infection