BMJ Open

The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-047575.R1
Article Type:	Communication
Date Submitted by the Author:	13-Jan-2021
Complete List of Authors:	Taylor, Julie; University College London, Clinical Operational Research Unit, Department of Mathematics Crowe, Sonya; University College London, Clinical Operational Research Unit, Department of Mathematics Espuny Pujol, Ferran; University College London, Clinical Operational Research Unit, Department of Mathematics Franklin, Rodney; Royal Brompton and Harefield NHS Foundation Trust Feltbower, Richard; University of Leeds, Paediatric Intensive Care Audit Network Norman, Lee; University of Leeds, Paediatric Intensive Care Audit Network Doidge, James; Intensive Care National Audit and Research Centre Gould, Doug; Intensive Care National Audit and Research Centre, Pagel, Christina; University College London, Clinical Operational Research Unit, Department of Mathematics
Primary Subject Heading:	Health services research
Secondary Subject Heading:	Research methods, Public health, Health informatics, Epidemiology
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Information management < BIOTECHNOLOGY & BIOINFORMATICS, EPIDEMIOLOGY, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, AUDIT

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement

Taylor JA[1], Crowe S[1], Espuny Pujol F[1], Franklin RC[2], Feltbower RG[3], Norman L[3], Doidge J[4], Gould D[4], Pagel C[1].

[1]Clinical Operational Research Unit, Department of Mathematics, University College London, UK

[2] Royal Brompton and Harefield NHS Foundation Trust, London, UK

[3]University of Leeds, Paediatric Intensive Care Audit Network, Leeds, West Yorkshire, UK

[4]Intensive Care National Audit and Research Centre, London, UK

Word Count: 2507 (3005 including box 1)

Corresponding Author:

Ms Julie Taylor, Clinical Operational Research Unit, Department of Mathematics, University College London, 4 Taviton Street, London.

Email: julie.a.taylor@ucl.ac.uk

The Opportunity

In the UK, challenges exist for many lifelong conditions such as congenital heart disease (CHD)[1], cancer[2], renal disease[3], haemophilia[4] and cystic fibrosis[5]. The UK has a wealth of high-quality registry and audit data.

We can improve healthcare services by better understanding current provision and outcomes[6] through linking datasets from clinical and national audits, registries, and other NHS activity data. In our research study, LAUNCHES QI (Box 1), we are using five linked national datasets to generate important understanding about services for CHD.

The Problem

Linking five national datasets has been challenging, laborious, and at times soul destroying. Coordinating applications to multiple data controllers and navigating the terminology, legal and governance structures can feel overwhelming. The process is time-consuming, complex and iterative.

Although a single study, much of the LAUNCHES process will be applicable to other studies, albeit with variations in university/institutional requirements, Research Ethics Committee (REC) requirements, the legal basis through General Data Protection Regulation (GDPR) compliance and the common law duty of confidentiality, and data controller specific requirements (Figure 1).

LAUNCHES QI: Linking AUdit and National datasets in Congenital HEart Services for Quality Improvement

The aim of LAUNCHES is to improve services for congenital heart disease and provide a template for other lifelong conditions by linking five national datasets to: describe patient trajectories through secondary and tertiary care; identify useful metrics for driving quality improvement (QI) and informing commissioning and policy; explore variation across services to identify priorities for QI.

Data Application Process

Planning the linkage process began with preparing the funding application. At that stage we liaised with each of the audits and developed an initial plan for linkage, including timelines and costs. This necessarily involved some guess work and we ended up underestimating both the time and costs involved.

The data application process in earnest (see figure 1) began in March 2018, with the start of the funding. The core dataset was the National Congenital Heart Disease Audit (NCHDA) dataset including cases from April 2000 to March 2017. Patients within NCHDA were matched to records in the Paediatric Intensive Care Audit network (PICANet), Intensive Care National Audit and Research Centre – Case Mix Programme (ICNARC-CMP), Hospital Episode Statistics (HES) and Office for National Statistics (ONS) mortality data. Resulting in a linked dataset allowing the analysis of longitudinal patient trajectories through the health care system. The data controller for NCHDA and PICANet is the Health Quality Improvement Partnership (HQIP), for ICNARC-CMP is ICNARC, and for HES and ONS is NHS Digital.

Our initial linkage plan was to use NHS Digital to act as a 'trusted third party', recommended as a standard approach to data linkage. However, following one of the author's experience and discussions with audit partners about risks of delay and feasibility, we opted for a method where each data controller linked their own data to the NCHDA identifiers.

Before beginning the process we met with colleagues attempting a similar study and were able to draw on their experience and utilise their draft forms, and our study coordinator had prior experience applying to NHS Digital, so we consider that we had a head start that potentially reduced the time taken to receive approvals. We had allowed six months to complete this process, and expected to have the linked five datasets ready for analysis shortly after this.

Required first were university approvals for: data protection, insurance, information governance, and approval of the completed Integrated Research Application System (IRAS) application, protocol and other study documents. This information and approvals made up the document set (Table 1) to submit to the Research Ethics Committee (REC) (required as LAUNCHES involved patients identified in the context of their past use of services in the NHS), and the Confidentiality Advisory Group to apply for section 251 approval to link and process the data (as it is not feasible to contact patients from the audits for retrospective individual consent). While these applications were under review, we could begin the four data applications for the five datasets, to submit for approval to the three data controllers.

Box 1 – Description of the LAUNCHES QI study and it's data application process.

Our Intention

We hope to help others applying for and linking registry, health administration and audit datasets by describing the processes, duration, resource requirements and challenges involved in LAUNCHES, and offering some tips and suggestions for system improvements based on our experience. We hope this paper will contribute to a renewed national conversation about balancing justifiably stringent requirements of data protection and governance with the potential opportunity for beneficial research capitalising on the wealth of healthcare data available.[7-13]

Challenges and learning

Volume of information required

Table 1 details the full set of documents for the data application processes (47 in total comprising 384 pages), which were required by 11 controllers or departments and submitted 161 times in total.

Although similar study information was requested for each of the seven application forms and the protocol, each required different wording, structure and detail. Several documents, designed for clinical trials and other prospective studies recruiting participants, were difficult to adapt to our study context.

	Regulator												
	Data protection*	Information Governance*	JRO office*	HRA REC	CAG	Named NHS Site	NHS site of 1 collaborator	HQIP NICOR	HQIP PICANet	ICNARC	NHS Digital	Pages/ screens	Times submitted
Study forms filled out													
Data Protection form	■		■		■							7	**3**
Governance forms (online)		■										1	**1**
IRAS form			■	■		■	■					32	**4**
Ethics amendment form 1			■	■		■	■					3	**4**
Ethics amendment form 2			■	■		■	■					4	**4**
CAG form – IRAS		■		■	■	■	■					40	**5**
CAT advice form					■							5	**1**
CAG amendment form 1					■	■	■				■	5	**4**
CAG amendment form 2					■	■	■			■		6	**4**
HQIP form including partner form				■		■		■				29	**3**
HQIP form (PICANet)									■			24	**1**
ICNARC form										■		5	**1**
DARS form (online)											■	21	**1**
Data fields & justification (NICOR)								■				4	**1**
Data fields & justification (PICANet)									■			4	**1**
Data fields & justification (ICNARC)										■		4	**1**
Data Linkage Diagram			■	■	■	■	■	■	■	■	■	1	**9**
Detailed data linkage diagram											■	1	**1**
Risk Assessment	■	■	■									3	**3**
IG toolkit confirmation				■	■	■	■	■	■	■	■	1	**8**
Correspondence with NICOR					■	■						1	**1**
Correspondence with charities					■							1	**1**
Letter from SIRO					■	■	■					1	**3**
Privacy notice			■	■	■	■	■	■	■	■	■	2	**9**
Protocol			■	■	■	■	■			■		25	**6**
Insurance registration form			■									6	**1**
Statement of Activities			■	■		■	■					22	**4**
Organisational Info Document			■	■		■	■					7	**4**
Cover Letter x 3				■	■							3	**2**
Checklist				■	■							2	**2**
Documents organised/shared													
HQIP letter of support			■	■	■							1	**3**
Insurance Certificate			■	■								1	**2**
REC provisional opinion			■		■							5	**2**
REC favourable opinion			■		■	■	■	■	■	■	■	4	**8**
Ethics amendment 1 approval			■		■	■	■				■	1	**5**
Ethics amendment 2 approval			■		■	■	■				■	1	**5**
CAG provisional approval			■	■		■	■					8	**4**
CAG approval			■	■		■	■	■	■	■	■	6	**8**
CAG amendment 1 approval			■	■		■	■				■	3	**5**
CAG amendment 2 approval			■	■		■	■			■	■	5	**6**
Funder Contract			■	■								38	**2**
PI CV			■	■		■	■					4	**4**
CV's from team						■						10	**1**
Reviewers Comments			■	■		■	■					4	**4**
HRA/HCRW approval letter			■			■	■	■	■	■	■	4	**7**
Training certificates						■						5	**1**
Service agreement						■						14	**1**
Total Documents Required	**2**	**3**	**25**	**22**	**19**	**32**	**23**	**8**	**8**	**11**	**13**	**384**	**161**

Table 1- the documents required for the study. *Block shading indicates documents required for each corresponding regulator. The 'times submitted' column sums the number of times each document was required by a different regulator. * Processes as at UCL. IRAS = Integrated Research Application System, CAG = Confidentiality Advisory Group, CAT = Confidentiality Advice Team, HQIP = Health Quality Improvement Partnership, HRA = Health Research Authority, PICANet = Paediatric Intensive Care Audit Network, ICNARC = Intensive Care National Audit & Research Centre, DARS = Data Access Request Service, NICOR = National Institute for Cardiovascular Outcomes Research, IG = Information Governance, SIRO = Senior Information Risk Owner, HQIP = Health Quality Improvement Partnership, HRA REC = Health Research Authority Research Ethics Committee, PI = Principal Investigator, CV = Curriculum Vitae, HCRW = Health and Care Research Wales, JRO = Joint Research Office.*

Requests for alterations and further information prior to approval were received between one and nine times for each data controller. Additional communications followed, for data preparation and transfer (Figure 2).

Following approvals, required further follow up on the study document set, e.g. chasing publication of privacy notices on partner websites, submitting amendments, annual reports to ethics and Confidentiality Advisory Group (CAG), and annual renewals for each data sharing agreement.

Person time required

It was tricky to assess in advance the person time required. Once the extent of the workload became clearer, we increased our 20% full time equivalent (FTE) study coordinator to 40% FTE for the beginning part of the study, but at the cost of less project coordinator support towards the end of the study (fortunately our funder was supportive). In retrospect the time required from the principal investigators to check, edit and input into the documentation was underestimated.

The teams working at the national audits also had limited resources and on occasions important day-to-day work necessarily took priority over processing data requests.

Costs

Allocating accurate funds for access to datasets is challenging. i) It is difficult to anticipate the final costs for some datasets and they were underestimated in the grant application. ii) Although you apply for data for the full length of the project, you are required to extend the data sharing agreement annually, this required additional funds for charges, study co-ordination time, and PI time. iii) In the case of NHS Digital, requesting even one additional field could result in a charge equal to that of the original data extract.

Process Delays

A timeline of the process is given in Figure 2. Study permissions, including university processes and receiving ethical and CAG approval, took eight months, and the data applications took between 3 and 7 months. Acquiring the data took a further 7-10 months.

Below we give examples of factors that caused delay and confusion to illustrate how seemingly small decisions had significant consequence and sometimes the complexity was overwhelming.

<u>One small decision…</u>

Small, seemingly inconsequential, decisions in filling out forms had far reaching consequences which did not become apparent until much later. In our case, a small decision

in the IRAS filter questions, made early to avoid redoing weeks of initial work on the IRAS form, ultimately ended up creating a lot of extra work. The resulting additional forms and required approvals are still being processed with an NHS site two years on.

Lack of clarity

We applied for two datasets (NCHDA, PICANet) that had the same data controller (HQIP), so we were unsure whether one or two applications were required. We were initially told one (which we prepared) but later told that two were needed. Therefore, two forms were submitted to the same data controller, to be discussed at the same meeting, with virtually identical content but slightly tailored to each specific dataset.

Requiring inside knowledge

Not all fields were listed in data dictionaries, and some derived fields that we asked for (e.g. age at an event instead of date) needed negotiation with each data provider. We were lucky to have project partners from each audit who knew that derived fields were possible.

Bureaucracy

Each time that we recruited research staff, some of the data sharing agreements required updating. Even such a minor change took up to 6 weeks and needed to be done separately for each data agreement, delaying when new staff could begin working with the data.

Being too specific

For IRAS and CAG we stated that we would be requesting life status and age at last known life status from ONS, but our advisory group subsequently suggested also requesting type of residence at death and hospital name. This change created a cascade of amendments resulting in significant delay. Had we stated that we would be applying for ONS data but not specified the fields, the amendment process would not have been necessary.

Not being specific enough

It was difficult to know how specific to be. As the data application developed, some changes to the fields were required and making these changes was a lengthy process. In our case, a failure to specify the number of decimal places required for age and our intention to request month and year of birth in our CAG application, coupled with a misunderstanding following asking for date of event in one dataset caused problems when it came to data extraction. This required a CAG amendment to resolve.

The impact of infrequent data controller meetings

In addition to the time spent preparing applications, timing applications to coincide with monthly (or sometimes less frequent) data controller approval meetings was difficult, and not all controllers provide information on meeting dates. Narrowly missing a meeting meant waiting a month or more for the next meeting, and if there were requests for further information then the process was delayed by another month. Additionally, an amendment for one data controller required changes to all other data applications, which lengthened the process considerably. We had five different approval bodies across the datasets.

Maintaining enthusiasm among project partners

One impact from process delays was the challenge of keeping project partners engaged and enthusiastic. The timeline our partners had originally agreed to had to be extended twice, with few interim developments to report back.

The fear of legal misstep, delaying sharing data

Failure to comply with the data protection principles could mean a fine of up to €20 million, or 4% of an organisation's total worldwide annual turnover, whichever is higher. Data controllers with good processes in place but without complete clarity on remit potentially contribute to delays because of concerns about a possible breach. For example, one data controller repeatedly asked us to clarify that our collaborators were advisory only without access to the linked data set, despite us providing that assurance at each stage.

Acquiring data

Once all approvals are in place, the process to gain access to and link the data is also lengthy. For example, following approval from the data controller (HQIP), it took seven months for the cohort to then be constructed and matching data sent to the other data providers for linkage.

Acquisition following approvals was delayed by a combination of: limited staff resources in NICOR for initiating linkage with other audits, amendments to obtain specific data fields, CAG amendments, and an amendment for a bespoke linkage plan (that was later shelved because of lack of data controller capacity). Whilst some of these delays could have been

avoided in hindsight, it is almost impossible to anticipate everything ahead of time, and the process for any changes is often as lengthy as a fresh application.

Top Tips

Love thy research assistants/study coordinators (the legal amount):

These are roles that often get squeezed at application stage to reduce costs and can be challenging to resource if part-time. But these roles are absolutely crucial in accessing data.

Find people who have been there before you:

Researchers who were further ahead in their data application journey shared their learning and expertise with us. This helped speed up the initial process of preparing the applications.

Have a flexible start up period:

Having an initial data application period in the project with just the Principal Investigators and a data coordinator funded allowed us to react early to delays and quickly adjust the timeline.

Add clinical and audit experts to the team:

Input from clinicians on data fields helped with data minimisation vs research realisation (particularly a problem for exploratory work). We also found it useful to include, where possible, collaborators working for data controllers and audits.

Plan your data minimisation and beware of differences in how some terms are interpreted:

You will need to describe how you have minimised the need for patient identifiers, and all fields requested will require justification. In doing so be sure to correctly state if data will be de-identified, pseudonymised or further anonymised. Some consider anonymised data to be aggregate data only, and amendments may be needed to clarify terms.

System improvement

It has taken us over 2.5 years to create our linked dataset, so most of our analyses and outputs on what was originally a two-year project are yet to be completed. But things could have been worse: we have an understanding funder and an incredible study coordinator, and we experienced no project staff turnover or major changes in audit staff.

Patient representatives on a different data linkage project that experienced similar challenges, expressed disbelief and dismay at how complicated the processes were. There is no centralised or easily accessible support for how to navigate the system – each new project team is on its own. So how might the system improve?

A shorter, streamlined system for retrospective data

Ethics & CAG

The forms for ethical and CAG approval are set up for clinical trials and cohorts prospectively recruiting participants. A shorter application form specific to routinely collected retrospective data would be helpful. CAG have now introduced their precedent set pathway[14] for an expedited review, which although a step in the right direction requires time-limited access to undertake record linkage/validation and anonymisation of data.

Single application

A template form with the same questions for all healthcare-related data controllers would reduce workload significantly. Each data controller would receive the same form, with the specification of all fields requested from each dataset.

Defined remit

The various application forms ask several questions regarding patient benefit (e.g. anticipated magnitude of patient benefit, how it will be achieved and measured, what contribution universities make to improving patient services, number of planned manuscripts), that are reiterated from funding, REC and CAG , but not consistently required in all data applications.

These questions were sometimes additional to the actual data application forms and so introduced delays as emails were sent back and forth (see figure 2). A joint system between funders, REC/CAG and data controllers so that each could be sure that important questions on feasibility, importance and process have been answered while minimising burden on applicants (and committees within each organisation) would be beneficial. In the absence of such a system, clarity on remit would reduce the pressure on the data controllers with the more conservative approaches, to obtain such detailed information.

Funders

Funders are increasingly acknowledging the time it will take to access and link data, and the uncertainties around timing. Additionally, they need to ensure that projects requiring external data have been resourced appropriately with sufficient budget for data access and project management. A more ambitious response could be centralised project coordinators within

funding bodies that are assigned to successful projects to help manage data access. This may be more efficient as the funders would build up expertise and experience of data access processes and make it easier to resource part-time project coordination.

Data controllers/processors; REC and CAG

Staff at the point of contact (on both sides) are bearing the brunt of the frustrated emails and the fraying patience of applicants working to their deadlines. Audits are working hard, encompassing much more than providing data to researchers, with their own staffing and resource issues, and have to abide by stringent data governance rules. The underlying systems seem to be the problem, but there are few opportunities for feedback.

Available guidance needs streamlining as it is currently lengthy, spread across a number of web pages, and often hard to follow or ambiguous.

Another possibility being adopted in other countries[15] is a central authority for processing and approving applications, and implementing linkage, thereby reducing the number of data controllers carrying out effectively the same review, and increasing efficiency.

Universities

Research assistants/study coordinators are really important for the success of this kind of research and play vital roles in information governance and data security, yet they are often funded on rolling temporary contracts with no job security or progression, often resulting in the loss of valuable expertise and experience. Universities should consider how to protect these roles, e.g. they could support permanent, centralised study coordinators who could

assist several projects across the university. Similar problems of job security also affect the data processors.

Tackling fear and confusion over data protection

How GDPR and other data protection laws apply to these sorts of retrospective studies is hard to understand and open to different interpretations. The importance of data governance and potential penalties for breaching GDPR mean that often the most conservative interpretation is used which often increases the complexity and workload. Clearer guidance, a single form and clarification on remit would alleviate this, as would guidance on the inclusion of collaborators and processes to share data after a dataset is established.

Conclusions

The current system is incredibly complex, arduous and slow, stifling innovation and delaying scientific progress. NHS data can inform and improve health services and we believe there is an ethical responsibility to use it to do so.[12]

Streamlining the number of applications required for accessing data for health services research and providing clarity to data controllers will facilitate the maintenance of stringent governance, while accelerating scientific studies and progress, leading to swifter application of findings and improvements in healthcare.

Acknowledgements

We would like to thank the data application teams at PICANet, ICNARC, NICOR, HQIP, and NHS Digital for their help and guidance as we negotiated the data application system.

Competing Interests

None declared.

Funding

This study is supported by the Health Foundation, an independent charity committed to bringing about better health and health care for people in the UK (Award number 685009).

References

1. Brown KL, Crowe S, Franklin R, et al. Trends in 30-day mortality rate and case mix for paediatric cardiac surgery in the UK between 2000 and 2010. *Open Heart* 2015;2(1):e000157. doi: 10.1136/openhrt-2014-000157 [published Online First: 2015/04/22]
2. Kowalczyk JR, Samardakiewicz M, Pritchard-Jones K, et al. European Survey on Standards of Care in paediatric oncology centres. *Eur J Cancer* 2016;61:11-9. doi: 10.1016/j.ejca.2016.03.073 [published Online First: 2016/05/01]
3. Plumb LA, Hamilton AJ, Inward CD, et al. Continually improving standards of care: The UK Renal Registry as a translational public health tool. *Pediatr Nephrol* 2018;33(3):373-80. doi: 10.1007/s00467-017-3688-2 [published Online First: 2017/06/24]
4. Dolan G, Makris M, Bolton-Maggs PH, et al. Enhancing haemophilia care through registries. *Haemophilia* 2014;20 Suppl 4:121-9. doi: 10.1111/hae.12406 [published Online First: 2014/04/26]
5. Stern M, Bertrand DP, Bignamini E, et al. European Cystic Fibrosis Society Standards of Care: Quality Management in cystic fibrosis. *J Cyst Fibros* 2014;13 Suppl 1:S43-59. doi: 10.1016/j.jcf.2014.03.011 [published Online First: 2014/05/27]
6. Deeny SR, Steventon A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ Qual Saf* 2015;24(8):505-15. doi: 10.1136/bmjqs-2015-004278 [published Online First: 2015/06/13]
7. Mourby MJ, Doidge J, Jones KH, et al. Health Data Linkage for UK Public Interest Research: Key Obstacles and Solutions. *Int J Popul Data Sci* 2019;4(1):1093. doi: 10.23889/ijpds.v4i1.1093 [published Online First: 2019/04/02]

8. Jones KH, Heys S, Tingay KS, et al. The Good, the Bad, the Clunky: Improving the Use of Administrative Data for Research. *Int J Popul Data Sci* 2018;4(1):587. doi: 10.23889/ijpds.v4i1.587 [published Online First: 2018/12/04]

9. Jones KH, Ford DV. Population data science: advancing the safe use of population data for public benefit. *Epidemiol Health* 2018;40:e2018061. doi: 10.4178/epih.e2018061 [published Online First: 2019/02/01]

10. Cavallaro FL-W, F; Cannings-John, R; Harron, K. Reducing barriers to data access for research in the public interest—lessons from covid-19. *BMJ Opinion* 2020

11. Dattani N, Hardelid P, Davey J, et al. Accessing electronic administrative health data for research takes time. *Arch Dis Child* 2013;98(5):391-2. doi: 10.1136/archdischild-2013-303730 [published Online First: 2013/03/02]

12. Jones KH, Laurie G, Stevens L, et al. The other side of the coin: Harm due to the non-use of health-related data. *Int J Med Inform* 2017;97:43-51. doi: 10.1016/j.ijmedinf.2016.09.010 [published Online First: 2016/12/07]

13. Lugg-Widger F AL, Cannings-John R, et al. . Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass. *International Journal of Population Data Science* 2018;3(3)

14. The CAG precedent set review pathway 2020 [updated 20/10/2020. Available from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/cag-precedent-set-review-pathway/ accessed 21/10/2020 2020.

15. Oderkirk J, Ronchi E, Klazinga N. International comparisons of health system performance among OECD countries: opportunities and data privacy protection challenges. *Health Policy* 2013;112(1-2):9-18. doi: 10.1016/j.healthpol.2013.06.006 [published Online First: 2013/07/23]

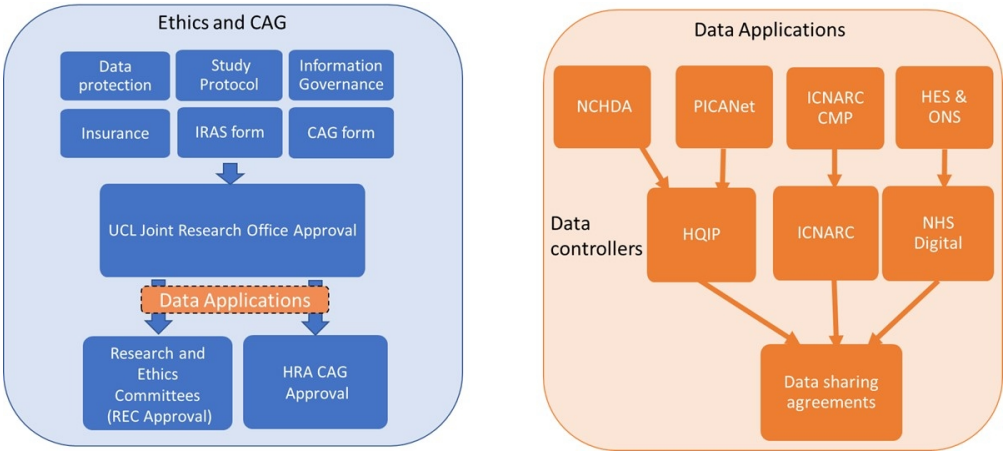

Figure 1 - Flow diagram of the study approval and data application process. CAG = Confidentiality Advisory Group, HRA = Health Research Authority, IRAS = Integrated Research Application System, HQIP = Health Quality Improvement Partnership, UCL = University College London, NCHDA = National Congenital Heart Disease Audit, PICANet = Paediatric Intensive Care Audit Network. ICNARC CMP = Intensive Care National Audit & Research Centre Case Mix Programme

203x91mm (150 x 150 DPI)

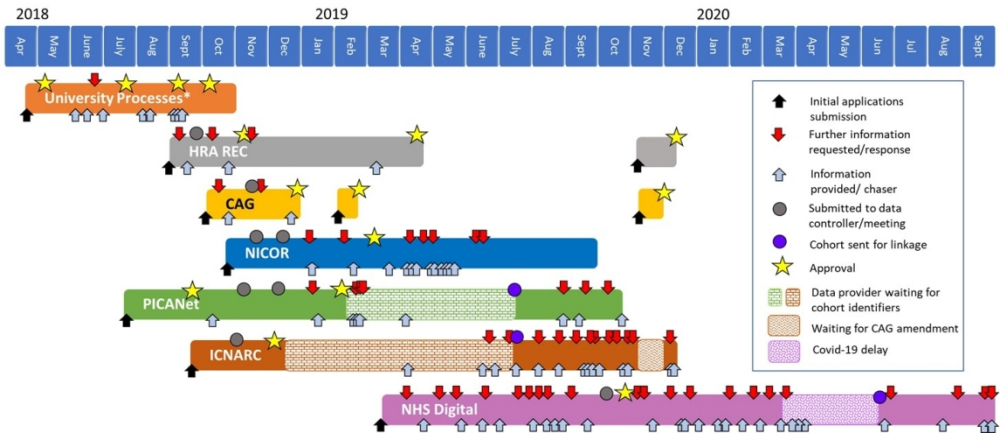

Figure 2 - Timeline of the permission process, order and time taken in LAUNCHES QI, from first submission to data acquisition. Each box represents the time from submission of the data application to transfer of the data to UCL, for each dataset. Note that not all requests for further information and responses have been included! *Processes as at UCL, HRA = Health Research Authority, REC = Research Ethics Committee, CAG = Confidentiality Advisory Group, NICOR =National Institute for Cardiovascular Outcomes Research, PICANet = Paediatric Intensive Care Audit Network, ICNARC = Intensive Care National Audit & Research Centre

246x107mm (150 x 150 DPI)

BMJ Open

The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-047575.R2
Article Type:	Communication
Date Submitted by the Author:	15-Jun-2021
Complete List of Authors:	Taylor, Julie; University College London, Clinical Operational Research Unit, Department of Mathematics Crowe, Sonya; University College London, Clinical Operational Research Unit, Department of Mathematics Espuny Pujol, Ferran; University College London, Clinical Operational Research Unit, Department of Mathematics Franklin, Rodney; Royal Brompton and Harefield NHS Foundation Trust Feltbower, Richard; University of Leeds, Paediatric Intensive Care Audit Network Norman, Lee; University of Leeds, Paediatric Intensive Care Audit Network Doidge, James; Intensive Care National Audit and Research Centre Gould, Doug; Intensive Care National Audit and Research Centre, Pagel, Christina; University College London, Clinical Operational Research Unit, Department of Mathematics
Primary Subject Heading:	Health services research
Secondary Subject Heading:	Research methods, Public health, Health informatics, Epidemiology
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Information management < BIOTECHNOLOGY & BIOINFORMATICS, EPIDEMIOLOGY, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, AUDIT

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement

Taylor JA[1], Crowe S[1], Espuny Pujol F[1], Franklin RC[2], Feltbower RG[3], Norman L[3], Doidge J[4], Gould D[4], Pagel C[1].

[1]Clinical Operational Research Unit, Department of Mathematics, University College London, UK

[2] Royal Brompton and Harefield NHS Foundation Trust, London, UK

[3]University of Leeds, Paediatric Intensive Care Audit Network, Leeds, West Yorkshire, UK

[4]Intensive Care National Audit and Research Centre, London, UK

Word Count excluding abstract: 3380 (3879 including box 1)

Corresponding Author:

Ms Julie Taylor, Clinical Operational Research Unit, Department of Mathematics, University College London, 4 Taviton Street, London.

Email: julie.a.taylor@ucl.ac.uk

Abstract

Background: We can improve health care services by better understanding current provision. One way to understand this is by linking datasets from clinical and national audits, national registries, and other NHS encounter data. However, getting to the point of having linked national datasets is challenging.

Objective: We describe our experience of the data application and linkage process for our study 'LAUNCHES QI', and the time, processes and resource requirements involved. To help others planning similar projects, we highlight challenges encountered and advice for applications in the current system as well as suggestions for system improvements.

Findings: The study set up for LAUNCHES QI began in March 2018, and the process through to data acquisition took 2.5 years.

Several challenges were encountered, including the amount of information required (often duplicate information in different formats across applications), lack of clarity on processes, resource constraints that limit an audit's capacity to fulfil requests, and the unexpected amount of time required from the study team. It is incredibly difficult to estimate the resources needed ahead of time, and yet necessary to do so as early on as funding applications. Early decisions can have a significant impact during latter stages and be hard to change, yet it is difficult to get specific information at the beginning of the process.

Conclusions: The current system is incredibly complex, arduous and slow, stifling innovation and delaying scientific progress. NHS data can inform and improve health services and we believe there is an ethical responsibility to use it to do so. Streamlining the number of applications required for accessing data for health services research and providing clarity to data controllers could facilitate the maintenance of stringent governance, while accelerating scientific studies and progress, leading to swifter application of findings and improvements in healthcare.

Limitations

- This is a single study using linked retrospective audit data, applicability of experiences and advice may vary according to the readers requirements.
- Clinical trials and studies recruiting participants will have a greater document burden than described here, but may also have found that the system was more applicable to their research than for retrospective data analysis.
- This datasets applied for covered English centres only as national data would have required several additional applications.

Strengths

- Provide valuable advice and insight to those embarking on research required linked national datasets, reassurance to those currently negotiating the system, and suggest improvements throughout the data application ecosystem for the future.
- Our experiences will overlap with those of prospective patient recruiting studies and clinical trials.

The Opportunity

In the UK, challenges exist for many lifelong conditions such as congenital heart disease (CHD)[1], cancer[2], renal disease[3], haemophilia[4] and cystic fibrosis[5]. The UK has a wealth of high-quality registry and audit data.

We can improve healthcare services by better understanding current provision and outcomes[6] through linking datasets from clinical and national audits, registries, and other NHS activity data. In our research study, LAUNCHES QI (Box 1), we are using five linked national datasets to generate important understanding about services for CHD.

The Problem

Linking five national datasets has been challenging, laborious, and at times demoralising and seemingly hopeless. Coordinating applications to multiple data controllers and navigating the terminology, legal and governance structures can feel overwhelming. The process is time-consuming, complex and iterative.

Although we are a single study, much of the LAUNCHES process will be applicable to other studies, albeit with variations in university/institutional requirements, Research Ethics Committee (REC) requirements, the legal basis through General Data Protection Regulation (GDPR) compliance and the common law duty of confidentiality, and data controller specific requirements (Figure 1).

LAUNCHES QI: Linking AUdit and National datasets in Congenital HEart Services for Quality Improvement

The aim of LAUNCHES is to improve services for congenital heart disease and provide a template for other lifelong conditions by linking five national datasets to: describe patient trajectories through secondary and tertiary care; identify useful metrics for driving quality improvement (QI) and informing commissioning and policy; explore variation across services to identify priorities for QI.

Data Application Process

Planning the linkage process began with preparing the funding application. At that stage we liaised with each of the audits and developed an initial plan for linkage, including timelines and costs. This necessarily involved some guess work and we ended up underestimating both the time and costs involved.

The data application process in earnest (see figure 1) began in March 2018, with the start of the funding. The core dataset was the National Congenital Heart Disease Audit (NCHDA) dataset including cases from April 2000 to March 2017. Patients within NCHDA were matched to records in the Paediatric Intensive Care Audit network (PICANet), Intensive Care National Audit and Research Centre – Case Mix Programme (ICNARC-CMP), Hospital Episode Statistics (HES) and Office for National Statistics (ONS) mortality data. Resulting in a linked dataset allowing the analysis of longitudinal patient trajectories through the health care system. The data controller for NCHDA and PICANet is the Health Quality Improvement Partnership (HQIP), for ICNARC-CMP is ICNARC, and for HES and ONS is NHS Digital.

Our initial linkage plan was to use NHS Digital to act as a 'trusted third party', recommended as a standard approach to data linkage. However, following one of the author's experience and discussions with audit partners about risks of delay and feasibility, we opted for a method where each data controller linked their own data to the NCHDA identifiers.

Before beginning the process we met with colleagues attempting a similar study and were able to draw on their experience and utilise their draft forms, and our study coordinator had prior experience applying to NHS Digital, so we consider that we had a head start that potentially reduced the time taken to receive approvals. We had allowed six months to complete this process, and expected to have the linked five datasets ready for analysis shortly after this.

Required first were university approvals for: data protection, insurance, information governance, and approval of the completed Integrated Research Application System (IRAS) application, protocol and other study documents. This information and approvals made up the document set (Table 1) to submit to the Research Ethics Committee (REC) (required as LAUNCHES involved patients identified in the context of their past use of services in the NHS), and the Confidentiality Advisory Group to apply for section 251 approval to link and process the data (as it is not feasible to contact patients from the audits for retrospective individual consent). While these applications were under review, we could begin the four data applications for the five datasets, to submit for approval to the three data controllers.

Box 1 – Description of the LAUNCHES QI study and it's data application process.

Our Intention

We hope to help others applying for and linking registry, health administration and audit datasets by describing the processes, duration, resource requirements and challenges involved in LAUNCHES, and offering some advice and suggestions for system improvements based on our experience. We hope this paper will contribute to a renewed national conversation about balancing justifiably stringent requirements of data protection and governance with the potential opportunity for beneficial research capitalising on the wealth of healthcare data available.[7-13]

Challenges and learning

Volume of information required

Table 1 details the full set of documents for the data application processes (47 in total comprising 384 pages), which were required by 11 controllers or departments and submitted 162 times in total.

Although similar study information was requested for each of the seven application forms and the protocol, each required different wording, structure and detail. Several documents, designed for clinical trials and other prospective studies recruiting participants, were difficult to adapt to our study context.

	Regulator												
	Data protection*	Information Governance*	JRO office*	HRA REC	CAG	Named NHS Site	NHS site of 1 collaborator	HQIP NICOR	HQIP PICANet	ICNARC	NHS Digital	Pages/ screens	Times submitted
Study forms filled out													
Data Protection form	■		■		■							7	**3**
Governance forms (online)		■										1	**1**
IRAS form			■	■		■	■					32	**4**
Ethics amendment form 1			■	■		■	■					3	**4**
Ethics amendment form 2			■	■		■	■					4	**4**
CAG form – IRAS		■		■	■	■	■					40	**5**
CAT advice form					■							5	**1**
CAG amendment form 1					■	■	■				■	5	**4**
CAG amendment form 2					■	■	■			■		6	**4**
HQIP form including partner form				■		■		■				29	**3**
HQIP form (PICANet)									■			24	**1**
ICNARC form										■		5	**1**
DARS form (online)											■	21	**1**
Data fields & justification (NICOR)								■				4	**1**
Data fields & justification (PICANet)									■			4	**1**
Data fields & justification (ICNARC)										■		4	**1**
Data Linkage Diagram			■	■	■	■	■	■	■	■	■	1	**9**
Detailed data linkage diagram											■	1	**1**
Risk Assessment	■	■	■									3	**3**
IG toolkit confirmation				■	■	■	■	■	■	■	■	1	**8**
Correspondence with NICOR					■	■						1	**2**
Correspondence with charities					■							1	**1**
Letter from SIRO					■	■	■					1	**3**
Privacy notice			■	■	■	■	■	■	■	■	■	2	**9**
Protocol			■	■	■	■	■			■		25	**6**
Insurance registration form			■									6	**1**
Statement of Activities			■	■		■	■					22	**4**
Organisational Info Document			■	■		■	■					7	**4**
Cover Letter x 3				■	■							3	**2**
Checklist				■	■							2	**2**
Documents organised/shared													
HQIP letter of support			■	■	■							1	**3**
Insurance Certificate			■	■								1	**2**
REC provisional opinion			■		■							5	**2**
REC favourable opinion			■		■	■	■	■	■	■	■	4	**8**
Ethics amendment 1 approval			■		■	■	■				■	1	**5**
Ethics amendment 2 approval			■		■	■	■				■	1	**5**
CAG provisional approval			■	■		■	■					8	**4**
CAG approval			■	■		■	■	■	■	■	■	6	**8**
CAG amendment 1 approval			■	■		■	■				■	3	**5**
CAG amendment 2 approval			■	■		■	■			■	■	5	**6**
Funder Contract			■	■								38	**2**
PI CV			■	■		■	■					4	**4**
CV's from team						■						10	**1**
Reviewers Comments			■	■		■	■					4	**4**
HRA/HCRW approval letter			■			■	■	■	■	■	■	4	**7**
Training certificates						■						5	**1**
Service agreement						■						14	**1**
Total Documents Required	**2**	**3**	**25**	**22**	**19**	**32**	**23**	**8**	**8**	**11**	**13**	**384**	**162**

Table 1- the documents required for the study. *Block shading indicates documents required for each corresponding regulator. The 'times submitted' column sums the number of times each document was required by a different regulator. * Processes as at UCL. IRAS = Integrated Research Application System, CAG = Confidentiality Advisory Group, CAT = Confidentiality Advice Team, HQIP = Health Quality Improvement Partnership, HRA = Health Research Authority, PICANet = Paediatric Intensive Care Audit Network, ICNARC = Intensive Care National Audit & Research Centre, DARS = Data Access Request Service, NICOR = National Institute for Cardiovascular Outcomes Research, IG = Information Governance, SIRO = Senior Information Risk Owner, HQIP = Health Quality Improvement Partnership, HRA REC = Health Research Authority Research Ethics Committee, PI = Principal Investigator, CV = Curriculum Vitae, HCRW = Health and Care Research Wales, JRO = Joint Research Office.*

Requests for alterations and further information prior to approval were received between one and nine times for each data controller. Additional communications followed, for data preparation and transfer (Figure 2).

Following approvals, required further follow up on the study document set, e.g. chasing publication of privacy notices on partner websites, submitting amendments, annual reports to ethics and Confidentiality Advisory Group (CAG), and annual renewals for each data sharing agreement.

Person time required

It was tricky to assess in advance the person time required. Once the extent of the workload became clearer, we increased our 20% full time equivalent (FTE) study coordinator to 40% FTE for the beginning part of the study, but at the cost of less project coordinator support towards the end of the study (fortunately our funder was supportive). In retrospect the time required from the principal investigators to check, edit and input into the documentation was underestimated.

The teams working at the national audits also had limited resources and on occasions important day-to-day work necessarily took priority over processing data requests.

Costs

Allocating accurate funds for access to datasets is challenging. i) It is difficult to anticipate the final costs for some datasets and they were underestimated in the grant application. ii) Although you apply for data for the full length of the project, you are required to extend the data sharing agreement annually, this required additional funds for charges, study co-ordination time, and PI time. iii) In the case of NHS Digital, requesting even one additional field after approval could result in a charge equal to that of the original data extract.

Process Delays

A timeline of the process is given in Figure 2. Study permissions, including university processes and receiving ethical and CAG approval, took eight months, and the data applications took between 3 and 7 months. Acquiring the data took a further 7-10 months.

Below we give examples of factors that caused delay and confusion to illustrate how seemingly small decisions had significant consequence and sometimes the complexity was overwhelming.

<u>One small decision…</u>

Small, seemingly inconsequential, decisions in filling out forms had far reaching consequences which did not become apparent until much later. In our case, a small decision

in the IRAS filter questions, made early to avoid redoing weeks of initial work on the IRAS form, ultimately ended up creating a lot of extra work. The resulting additional forms and required approvals are still being processed with an NHS site two years on.

Lack of clarity

We applied for two datasets (NCHDA, PICANet) that had the same data controller (HQIP), so we were unsure whether one or two applications were required. We were initially told one (which we prepared) but later told that two were needed. Therefore, two forms were submitted to the same data controller, to be discussed at the same meeting, with virtually identical content but slightly tailored to each specific dataset.

Requiring inside knowledge

Not all fields were listed in data dictionaries, and some derived fields that we asked for (e.g. age at an event instead of date) needed negotiation with each data provider. We were lucky to have project partners from each audit who knew that derived fields were possible.

Bureaucracy

Each time that we recruited research staff, some of the data sharing agreements required updating. Even such a minor change took up to 6 weeks and needed to be done separately for each data agreement, delaying when new staff could begin working with the data.

Being too specific

For IRAS and CAG we stated that we would be requesting life status and age at last known life status from ONS, but our advisory group subsequently suggested also requesting type of residence at death and hospital name. This change created a cascade of amendments resulting in significant delay. Had we stated that we would be applying for ONS data but not specified the fields, the amendment process would not have been necessary.

Not being specific enough

It was difficult to know how specific to be. As the data application developed, some changes to the fields were required and making these changes was a lengthy process. In our case, a failure to specify the number of decimal places required for age and our intention to request month and year of birth in our CAG application, coupled with a misunderstanding following asking for date of event in one dataset caused problems when it came to data extraction. This required a CAG amendment to resolve.

The impact of infrequent data controller meetings

In addition to the time spent preparing applications, timing applications to coincide with monthly (or sometimes less frequent) data controller approval meetings was difficult, and not all controllers provide information on meeting dates. Narrowly missing a meeting meant waiting a month or more for the next meeting, and if there were requests for further information then the process was delayed by another month. Additionally, an amendment for one data controller required changes to all other data applications, which lengthened the process considerably. We had five different approval bodies across the datasets.

Maintaining enthusiasm among project partners

One impact from process delays was the challenge of keeping project partners engaged and enthusiastic. The timeline our partners had originally agreed to had to be extended twice, with few interim developments to report back.

The fear of legal misstep, delaying sharing data

Failure to comply with the data protection principles could mean a fine of up to €20 million, or 4% of an organisation's total worldwide annual turnover, whichever is higher[14]. Data controllers with good processes in place but without complete clarity on remit potentially contribute to delays because of concerns about a possible breach. For example, one data controller repeatedly asked us to clarify that our collaborators were advisory only without access to the linked data set, despite us providing that assurance at each stage.

Acquiring data

Once all approvals are in place, the process to gain access to and link the data is also lengthy. For example, following approval from the data controller (HQIP), it took seven months for the cohort to then be constructed and matching data sent to the other data providers for linkage.

Acquisition following approvals was delayed by a combination of: limited staff resources in NICOR for initiating linkage with other audits, amendments to obtain specific data fields, CAG amendments, and an amendment for a bespoke linkage plan (that was later shelved because of lack of data controller capacity). Whilst some of these delays could have been

avoided in hindsight, it is almost impossible to anticipate everything ahead of time, and the process for any changes is often as lengthy as a fresh application.

Reusing the data

Having completed the data linkage, funding was received for another study that would use the same dataset. For most of the datasets, entirely new applications were required to use the existing data for this new but similar purpose, and despite our previous experience in LAUNCHES, it has taken another two years to complete the process, with the NHS Digital application still pending.

Useful advice

Love thy research assistants/study coordinators (the legal amount):

These are roles that often get squeezed at application stage to reduce costs and can be challenging to resource if part-time. But these roles are absolutely crucial in accessing data.

Find people who have been there before you:

Researchers who were further ahead in their data application journey shared their learning and expertise with us. This helped speed up the initial process of preparing the applications.

Have a flexible start up period:

Having an initial data application period in the project with just the Principal Investigators and a data coordinator funded allowed us to react early to delays and quickly adjust the timeline.

Add clinical and audit experts to the team:

Input from clinicians on data fields helped with data minimisation vs research realisation (particularly a problem for exploratory work). We also found it useful to include, where possible, collaborators working for data controllers and audits.

Plan your data minimisation and beware of differences in how some terms are interpreted:

You will need to describe how you have minimised the need for patient identifiers, and all fields requested will require justification. In doing so be sure to correctly state if data will be de-identified, pseudonymised or further anonymised. Some consider anonymised data to be aggregate data only, and amendments may be needed to clarify terms.

System improvement

It has taken us over 2.5 years to create our linked dataset, so most of our analyses and outputs on what was originally a two-year project are yet to be completed. But things could have been worse: we have an understanding funder and an incredible study coordinator, and we experienced no project staff turnover or major changes in audit staff.

There is no centralised or easily accessible support for how to navigate the system – each new project team is on its own. So how might the system improve? Below are our suggestions, which are also summarised in figure 3.

A shorter, streamlined system for retrospective data

Ethics & CAG

The forms for ethical and CAG approval are set up for clinical trials and cohorts prospectively recruiting participants. A shorter application form specific to routinely collected

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

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

Single application

A template form with the same questions for all healthcare-related data controllers would reduce workload significantly. Each data controller would receive the same form, with the specification of all fields requested from each dataset.

Defined remit

The various application forms ask several questions regarding patient benefit (e.g. anticipated magnitude of patient benefit, how it will be achieved and measured, what contribution universities make to improving patient services, number of planned manuscripts), that are reiterated from funding, REC and CAG , but not consistently required in all data applications.

These questions were sometimes additional to the actual data application forms and so introduced delays as emails were sent back and forth (see figure 2). A joint system between funders, REC/CAG and data controllers so that each could be sure that important questions on feasibility, importance and process have been answered while minimising burden on applicants (and committees within each organisation) would be beneficial. In the absence of such a system, clarity on remit would reduce the pressure on the data controllers with the more conservative approaches to obtain such detailed information, and in some cases prevent duplication of review of ethics, CAG applications, amendments and annual reports. For context, some applications required details of research outputs and benefits. The amendment to use the same data for another study (see 'reusing data') resulted in the re-review of the initial application with further questions on this already approved section. The questions referred to whether some benefits listed would actually be considered outputs. Defining remit could consider the level of detail necessary for each part of the process.

Funders

Funders are increasingly acknowledging the time it will take to access and link data, and the uncertainties around timing. Additionally, they need to ensure that projects requiring external data have been resourced appropriately with sufficient budget for data access and project management. A more ambitious response could be centralised project coordinators within

funding bodies that are assigned to successful projects to help manage data access. This may be more efficient as the funders would build up expertise and experience of data access processes and make it easier to resource part-time project coordination.

Data controllers/processors; REC and CAG

Staff at the point of contact (on both sides) are bearing the brunt of the frustrated emails and the fraying patience of applicants working to their deadlines. Audits are working hard, encompassing much more than providing data to researchers, with their own staffing and resource issues, and have to abide by stringent data governance rules. The underlying systems seem to be the problem, but there are few opportunities for feedback.

Available guidance needs streamlining as it is currently lengthy, spread across a number of web pages, and often hard to follow or ambiguous. Some independent organisations have attempted to provide guidance to help navigate the system. For example the MRC have developed an online tool to help determine the approvals needed,[16] and HQIP have just funded and published short films and written resources on understanding health data access designed to improve accessible, introductory information about the rules and processes governing access to health data.[17]

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

Universities

Research assistants/study coordinators are really important for the success of this kind of research and play vital roles in information governance and data security, yet they are often funded on rolling temporary contracts with no job security or progression, often resulting in the loss of valuable expertise and experience. Universities should consider how to protect these roles, e.g. they could support permanent, centralised study coordinators who could assist several projects across the university. Similar problems of job security also affect the data processors.

We also experienced an increase in the requirements of the Research and Development ("R&D") departments of the Universities and the NHS sites, with little if any flexibility to take study type into account.

Tackling fear and confusion over data protection

How GDPR and other data protection laws apply to these sorts of retrospective studies is hard to understand and open to different interpretations. The importance of data governance and potential penalties for breaching GDPR mean that often the most conservative interpretation is used which often increases the complexity and workload. Clearer guidance, a single form and clarification on remit would alleviate this, as would guidance on the inclusion of collaborators and processes to share data after a dataset is established.

Lessons from the pandemic

During 2020, research governance was relaxed in pursuit of rapid scientific evidence into COVID-19 aetiology, risk factors, and treatments. This included a fast track review for ethics review,[22] the pause of the need for approval under Regulation 3(4) of the Health Service Control of Patient Information Regulations 2002,[23] and the release of data and change in process by some data controllers.[24] This expedited progress of all studies into COVID, and

prioritisation of COVID research, shows that there is room to simplify the process.[10] This provides an important opportunity to learn from what happened during the COVID pandemic, and what could be adapted for the future.

Strengths & Limitations

Our experience of an ambitious project to link five datasets may not be entirely generalisable to other researchers, especially those applying for fully anonymised data from an established research database.

However for most studies, although some differences would apply for ethics and CAG requirements, such requirements will still need to be considered at the outset of the study. Datasets required will differ for different projects, but we expect that each research team will encounter similar governance requirements albeit with different requirements of the relevant data controllers. Longitudinal studies and clinical trials may have many approvals in place, but may be still experiencing their own challenges with changing governance requirements that could delay renewal of permissions.

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

Conclusions

The current system is incredibly complex, arduous and slow, stifling innovation and delaying scientific progress. NHS data can inform and improve health services and we believe there is an ethical responsibility to use it to do so.[12]

Streamlining the number of applications required for accessing data for health services research and providing clarity to data controllers will facilitate the maintenance of stringent governance, while accelerating scientific studies and progress, leading to swifter application of findings and improvements in healthcare.

Author Contributions

CP, SC and JT were involved in conceptualization, design and writing of the manuscript, JT wrote the first draft. CP, SC, JT, FEP, RCF, RGF, LN, JD, and DG were involved in the review of the manuscript.

Acknowledgements

We would like to thank the data application teams at PICANet, ICNARC, NICOR, HQIP, and NHS Digital for their help and guidance as we negotiated the data application system.

Competing Interests

None declared.

Funding

This study is supported by the Health Foundation, an independent charity committed to bringing about better health and health care for people in the UK (Award number 685009).

Figure Captions

Figure 1 - Flow diagram of the study approval and data application process. CAG = Confidentiality Advisory Group, HRA = Health Research Authority, IRAS = Integrated Research Application System, HQIP = Health Quality Improvement Partnership, UCL = University College London, NCHDA = National Congenital Heart Disease Audit, PICANet = Paediatric Intensive Care Audit Network. ICNARC CMP = Intensive Care National Audit & Research Centre Case Mix Programme.

Figure 2 - Timeline of the permission process, order and time taken in LAUNCHES QI, from first submission to data acquisition. Each box represents the time from submission of the data application to transfer of the data to UCL, for each dataset. Note that not all requests for further information and responses have been included! *Processes as at UCL, HRA = Health Research Authority, REC = Research Ethics Committee, CAG = Confidentiality Advisory Group, NICOR =National Institute for Cardiovascular Outcomes Research, PICANet = Paediatric Intensive Care Audit Network, ICNARC = Intensive Care National Audit & Research Centre.

Figure 3 – Illustrates the suggested system improvements for the future. The coloured bubbles contain the suggestions, with each larger bubble representing the authority that should consider these changes. CAG=Confidentiality Advisory Group.

References

1. Brown KL, Crowe S, Franklin R, et al. Trends in 30-day mortality rate and case mix for paediatric cardiac surgery in the UK between 2000 and 2010. *Open Heart* 2015;2(1):e000157. doi: 10.1136/openhrt-2014-000157 [published Online First: 2015/04/22]

2. Kowalczyk JR, Samardakiewicz M, Pritchard-Jones K, et al. European Survey on Standards of Care in paediatric oncology centres. *Eur J Cancer* 2016;61:11-9. doi: 10.1016/j.ejca.2016.03.073 [published Online First: 2016/05/01]

3. Plumb LA, Hamilton AJ, Inward CD, et al. Continually improving standards of care: The UK Renal Registry as a translational public health tool. *Pediatr Nephrol* 2018;33(3):373-80. doi: 10.1007/s00467-017-3688-2 [published Online First: 2017/06/24]

4. Dolan G, Makris M, Bolton-Maggs PH, et al. Enhancing haemophilia care through registries. *Haemophilia* 2014;20 Suppl 4:121-9. doi: 10.1111/hae.12406 [published Online First: 2014/04/26]

5. Stern M, Bertrand DP, Bignamini E, et al. European Cystic Fibrosis Society Standards of Care: Quality Management in cystic fibrosis. *J Cyst Fibros* 2014;13 Suppl 1:S43-59. doi: 10.1016/j.jcf.2014.03.011 [published Online First: 2014/05/27]

6. Deeny SR, Steventon A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ Qual Saf* 2015;24(8):505-15. doi: 10.1136/bmjqs-2015-004278 [published Online First: 2015/06/13]

7. Mourby MJ, Doidge J, Jones KH, et al. Health Data Linkage for UK Public Interest Research: Key Obstacles and Solutions. *Int J Popul Data Sci* 2019;4(1):1093. doi: 10.23889/ijpds.v4i1.1093 [published Online First: 2019/04/02]

8. Jones KH, Heys S, Tingay KS, et al. The Good, the Bad, the Clunky: Improving the Use of Administrative Data for Research. *Int J Popul Data Sci* 2018;4(1):587. doi: 10.23889/ijpds.v4i1.587 [published Online First: 2018/12/04]

9. Jones KH, Ford DV. Population data science: advancing the safe use of population data for public benefit. *Epidemiol Health* 2018;40:e2018061. doi: 10.4178/epih.e2018061 [published Online First: 2019/02/01]

10. Cavallaro FL-W, F; Cannings-John, R; Harron, K. Reducing barriers to data access for research in the public interest—lessons from covid-19. *BMJ Opinion* 2020

11. Dattani N, Hardelid P, Davey J, et al. Accessing electronic administrative health data for research takes time. *Arch Dis Child* 2013;98(5):391-2. doi: 10.1136/archdischild-2013-303730 [published Online First: 2013/03/02]

12. Jones KH, Laurie G, Stevens L, et al. The other side of the coin: Harm due to the non-use of health-related data. *Int J Med Inform* 2017;97:43-51. doi: 10.1016/j.ijmedinf.2016.09.010 [published Online First: 2016/12/07]

13. Lugg-Widger F AL, Cannings-John R, et al. . Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: Managing the morass. *International Journal of Population Data Science* 2018;3(3)

14. Data Protection Act: Part 6 Enforcement; Penalties. Section 155-159. 2018 [Available from: https://www.legislation.gov.uk/ukpga/2018/12/part/6/crossheading/penalties/enacted

15. The CAG precedent set review pathway 2020 [updated 20/10/2020. Available from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/cag-precedent-set-review-pathway/ accessed 21/10/2020 2020.

16. UKRI MRC Regulatory Support Centre. Health Data Access Tool Kit. [Available from: https://hda-toolkit.org/story_html5.html accessed 01/06/2021.

17. Health Quality Improvement Partnership. Understanding Health Data Access (UHDA). [Online guidance]. Available from: https://www.hqip.org.uk/understanding-health-data-access/#.YKuVhKhKg2w accessed 01/06/2021.

18. Oderkirk J, Ronchi E, Klazinga N. International comparisons of health system performance among OECD countries: opportunities and data privacy protection challenges. *Health Policy* 2013;112(1-2):9-18. doi: 10.1016/j.healthpol.2013.06.006 [published Online First: 2013/07/23]

19. O'Reilly D, Bateson O, McGreevy G, et al. Administrative Data Research Northern Ireland (ADR NI). *Int J Popul Data Sci* 2020;4(2):1148. doi: 10.23889/ijpds.v4i2.1148 [published Online First: 2020/09/17]

20. Nelson E, Burns F. Impact through engagement: Co-production of administrative data research and the approach of the Administrative Data Research Centre Northern Ireland. *Int J Popul Data Sci* 2020;5(3):1369. doi: 10.23889/ijpds.v5i3.1369 [published Online First: 2021/05/20]

21. Van Der Wel KA, Ostergren O, Lundberg O, et al. A gold mine, but still no Klondike: Nordic register data in health inequalities research. *Scand J Public Health* 2019;47(6):618-30. doi: 10.1177/1403494819858046 [published Online First: 2019/07/12]

22. NHS Health Research Authority. Fast track review guidance for COVID-19 studies. [updated 20/05/2021. Available from: https://www.hra.nhs.uk/covid-19-research/fast-track-review-guidance-covid-19-studies/ accessed 01/06/2021.

23. Department of Health and Social Care. Coronavirus (COVID-19): notification to organisations to share information., 2020. https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information

24. NHS Digital. DARS coronavirus (COVID-19) response [updated 26/05/2021. Available from: https://digital.nhs.uk/services/data-access-request-service-dars accessed 01/06/2021.

25. Macnair A, Love SB, Murray ML, et al. Accessing routinely collected health data to improve clinical trials: recent experience of access. *Trials* 2021;22(1):340. doi: 10.1186/s13063-021-05295-5 [published Online First: 2021/05/12]

26. Gilbert R, Lafferty R, Hagger-Johnson G, et al. GUILD: GUidance for Information about Linking Data sets. *J Public Health (Oxf)* 2018;40(1):191-98. doi: 10.1093/pubmed/fdx037 [published Online First: 2017/04/04]

27. Rivera DR, Gokhale MN, Reynolds MW, et al. Linking electronic health data in pharmacoepidemiology: Appropriateness and feasibility. *Pharmacoepidemiol Drug Saf* 2020;29(1):18-29. doi: 10.1002/pds.4918 [published Online First: 2020/01/18]

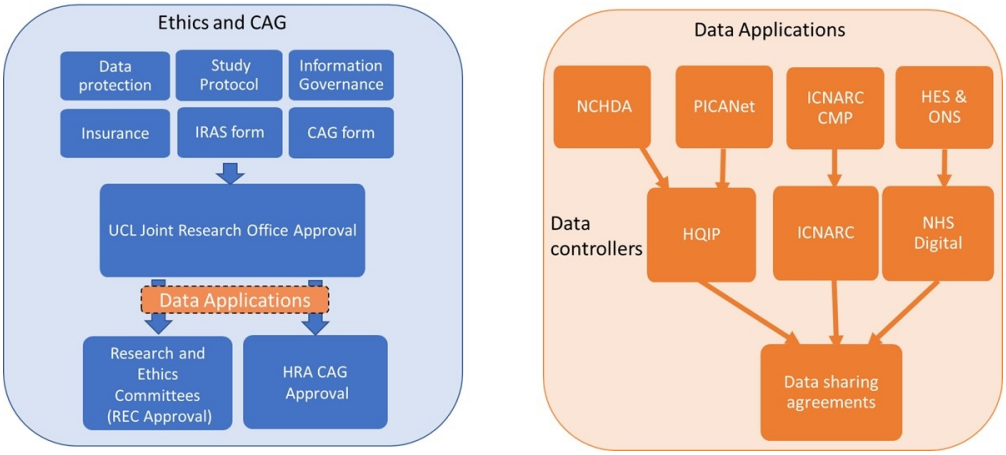

Figure 1 - Flow diagram of the study approval and data application process. CAG = Confidentiality Advisory Group, HRA = Health Research Authority, IRAS = Integrated Research Application System, HQIP = Health Quality Improvement Partnership, UCL = University College London, NCHDA = National Congenital Heart Disease Audit, PICANet = Paediatric Intensive Care Audit Network. ICNARC CMP = Intensive Care National Audit & Research Centre Case Mix Programme

203x91mm (150 x 150 DPI)

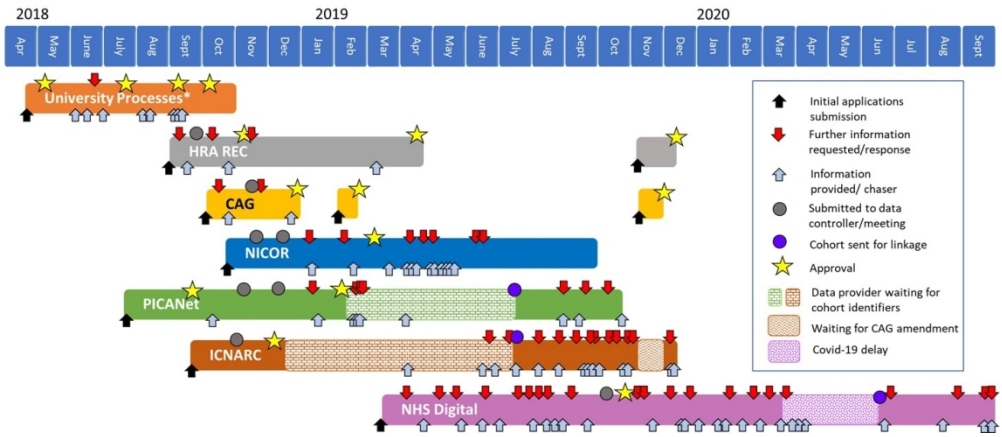

Figure 2 - Timeline of the permission process, order and time taken in LAUNCHES QI, from first submission to data acquisition. Each box represents the time from submission of the data application to transfer of the data to UCL, for each dataset. Note that not all requests for further information and responses have been included! *Processes as at UCL, HRA = Health Research Authority, REC = Research Ethics Committee, CAG = Confidentiality Advisory Group, NICOR =National Institute for Cardiovascular Outcomes Research, PICANet = Paediatric Intensive Care Audit Network, ICNARC = Intensive Care National Audit & Research Centre

246x107mm (150 x 150 DPI)

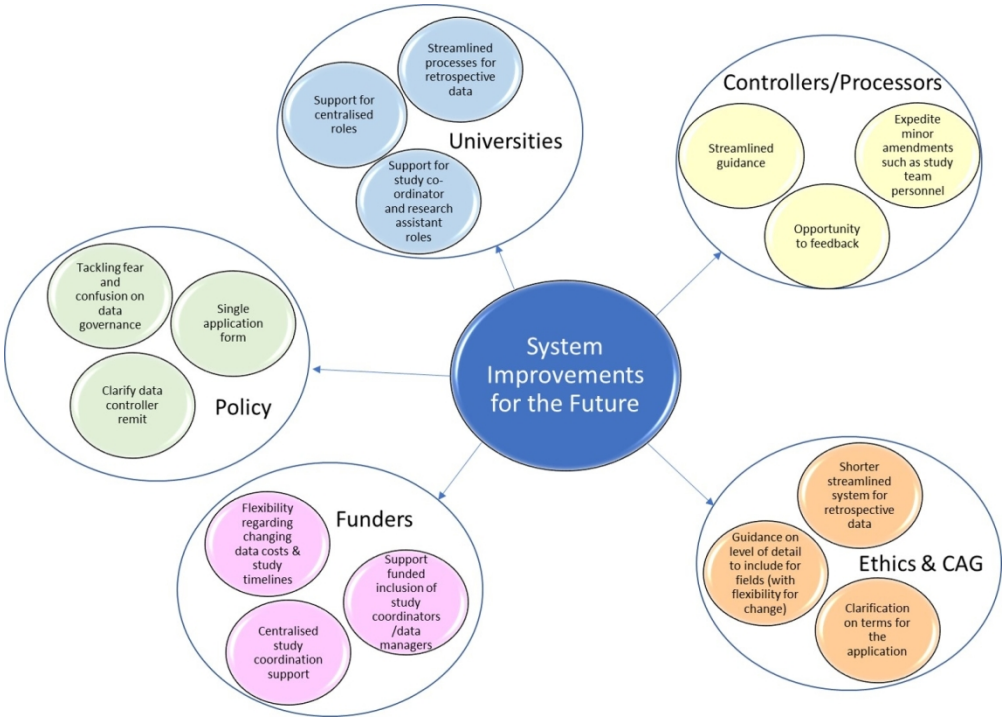

Figure 3 – Illustrates the suggested system improvements for the future. The coloured bubbles contain the suggestions, with each larger bubble representing the authority that should consider these changes. CAG=Confidentiality Advisory Group

267x190mm (149 x 149 DPI)