HDRUK
Health Data Research UK

Population Research UK

Design and Dialogue consultation outputs

09/07/2021

PRUK stakeholder consultation: summary of activity

In February and April 2021, HDR UK conducted a programme of consultation to inform the design and function of Population Research UK.
The purpose was to:

- Understand current experience, what works well and challenges in the current UK LPS research ecosystem that reduce the efficiency and effectiveness with which LPS are utilised to deliver useful insights.
- Define the roles that PRUK could play to reduce or eliminate these obstacles and improve the utilisation of LPS in the UK

The consultation included:

- A programme of 67 interviews. Interviewees had experiences of both running LPS and using LPS data and were selected from within and outside academia. Interviews included people in early, mid and senior stages of their careers, and people whose interest is primarily in the social sciences (27) (SS) and health sciences (40) (HS).
- An online survey. There were 216 respondents and 107 complete survey responses. Respondents were predominantly UK academic audience that was broad across geography, numbers of years experience in working with of Longitudinal Population Studies (LPS) and disciplinary background.

PRUK stakeholder consultation: summary of current experiences and obstacles

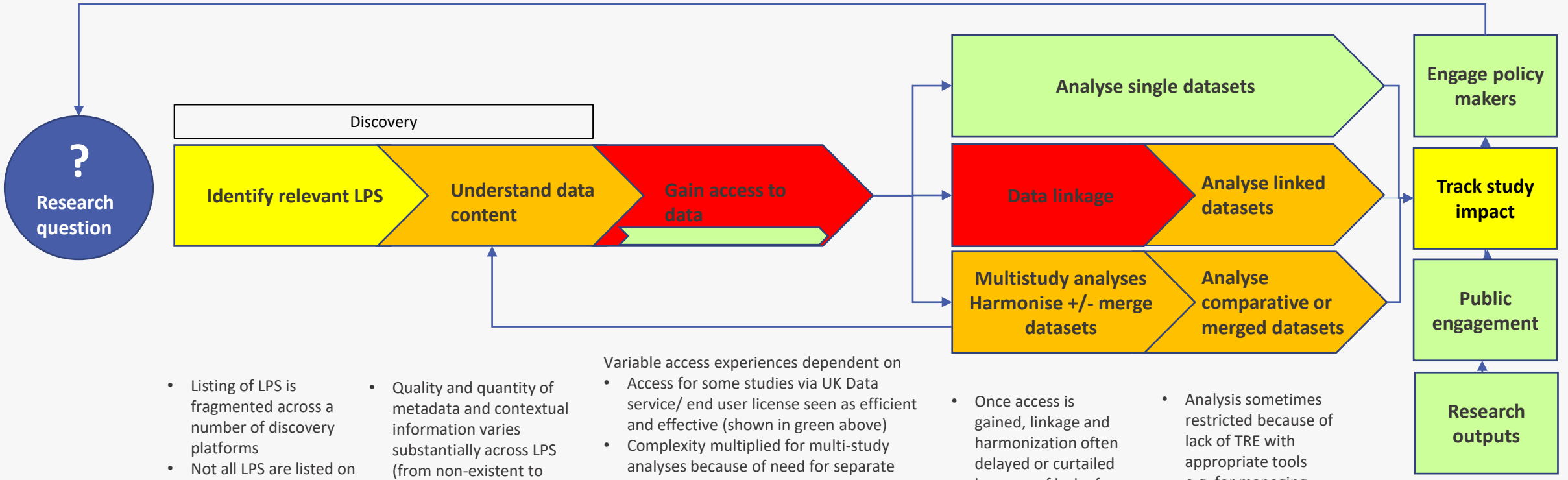
- Respondents were asked about their current experiences of using and working with LPS; the focus of the consultation was on the steps by which LPS data was made available, accessed, distributed and used for research. However, respondents also discussed other areas including the design of studies, data collection, training and stakeholder engagement.
- LPS Discovery and access is variable across datasets and disciplines. Currently challenges with understanding contents of datasets and getting access to data can cause studies to be delayed. There are multiple initiatives that support discovery of LPS, and the was a range of views on the level of additional activity and investment needed in these areas. There is a need for both high study-level discovery and both variable-level, as well as tools that enable greater understanding of dataset contents before a data access request.
- Data access was regarded as highly variable between social and biomedical LPS, and access was particularly problematic as the number of datasets requested increased. Streamlining access, and reducing required documentation were as important to the efficiency and effectiveness of LPS.
- Creating and accessing datasets linked to health and administrative data was the most consistent obstacle many respondents identified regardless of discipline. A high proportion of respondents would like to work more with LPS linked to a broad range of health and administrative datasets. The most common barriers cited for not working with linked data were difficulties with access, and lack of awareness of data available.
- Another theme of respondents addressed tackling underlying issues in the field, such as the representativeness of current studies of the general population, incentives for data sharing, underfunding of studies and the need for analytical methods skills and training to support complex data analysis.

PRUK stakeholder consultation: current experience and obstacles

HDRUK
Health Data Research UK

Obstacles mean that analyses are excessively delayed, curtailed or even abandoned

Highly efficient and effective



- Listing of LPS is fragmented across a number of discovery platforms
- Not all LPS are listed on discovery platforms
- However, generally considered a surmountable problem

- Quality and quantity of metadata and contextual information varies substantially across LPS (from non-existent to comprehensive metadata)
- Lack of resource and expertise in some LPS to develop metadata
- Some tools for querying data exist but are not widespread

Variable access experiences dependent on

- Access for some studies via UK Data service/ end user license seen as efficient and effective (shown in green above)
- Complexity multiplied for multi-study analyses because of need for separate approval for each study.
- Variable practice across biomedical cohorts

For others –

- Access processes often idiosyncratic and not transparent
- Excessive documentation often required
- Some LPS require collaboration with PIs in order to gain access
- Delay in approval often arises from lack of resource
- Cost

- Once access is gained, linkage and harmonization often delayed or curtailed because of lack of resources

- Analysis sometimes restricted because of lack of TRE with appropriate tools e.g. for managing large 'omics datasets
- There is a needed for expanding knowledge and expertise in analysis of complex datasets

Source:PRUK design and dialogue interviews and questionnaire, February – March 2021

PRUK stakeholder consultation: potential roles for PRUK

Respondents identified several roles PRUK could play in advancing the field.These included:

- Developing standards and process models for use by studies (e.g. metadata and discovery standards, access standards)
- Provision of resources (funding, capacity and expertise) to enable LPS to carry out operations that are currently underfunded
- Providing data services; conducting data linkage, and to a lesser extent dataset harmonisation, and provision of the resulting datasets to researchers
- Provision of infrastructure such as a discovery platforms, data access platforms and trusted research environments (TREs) to support analysis
- Convening the community, though creating forum for the sharing of best practice and expertise

- Some respondents PRUK should provide an all-encompassing solution, including a comprehensive, searchable discovery platform with rich metadata covering all LPS; a streamlined, multi-cohort access process; development and provision of linked and merged datasets; and bespoke TREs for LPS analysis. Others suggested a role of joining up existing initiatives.
- Several interviewees also saw a role for PRUK in closing gaps in policy and processes between biomedical studies and social science studies in order to broaden the usage of LPS. Several others suggested that PRUK should have more of a role in promoting an inclusive research agenda, recognising that some groups and demographics were currently underrepresented in LPS research.
- The survey outputs support the direction of travel of PRUK with a focus on streamlining data access and data linkage. In developing PRUK, respondents highlighted the need for PRUK to work with and build on good practice. There is a need for initiatives to be more joined up and interoperable. The consultation also highlighted underlying tension that at a national level there must be the right balance between investment in LPS to collect, manage and share data and meta-initiatives such as PRUK.

Summary: suggested roles for PRUK

HDRUK
Health Data Research UK

Standard setting

Set key standards for LPS, e.g.,

- Data standards
- Metadata standards
- Metadata and data in approved formats
- Listing on agreed discovery platforms
- Approval process conforming to approved model
- Specify the provisions of a TRE
- Archiving standard for completed studies

Funding

Provide funds and/or PRUK resource, for example to

- Enable studies (including studies no longer actively funding/collecting data) to achieve standards (e.g., developing metadata)
- Enable studies to provide an efficient access process
- Pilot harmonized approaches to data collection across LPS

Data services

Developing desired datasets for users:

- Creating and providing linked datasets
- Creating and providing merged datasets
- Providing data in consistent formats and common data models
- Harmonising variables
- Creating and providing merged datasets

Infrastructure provision

Develop and maintain (in-house or outsourced) LPS-related infrastructure, including some or all of:

- Portal to 3rd party discovery platforms
- Comprehensive LPS discovery platform
- Trusted research environments
- Repository of tools and algorithms
- Training programmes

Convene the community

Creating forum/ consortia for the sharing of best practice and expertise

- Working groups of interest to individual LPS
- Providing a collective perspective to stakeholders e.g. research funders.
- Policy engagement
- Public engagement

PRUK could play any combination of these five roles to solve the key problems in utilization of LPS

Interview findings

PRUK design and dialogue development programme stakeholder interviews

- In February and March 2021, a programme of 67 interviews was undertaken as part of the design phase for Population Research UK (PRUK)
- Each interview was a 45–60-minute, semi-structured interview that sought to:
  - Elucidate the obstacles in the current UK research ecosystem that reduce the efficiency and effectiveness with which longitudinal population studies (LPS) are utilised to deliver useful insights
  - Define the roles that PRUK could play to reduce or eliminate these obstacles and thus improve the utilisation of LPS in the UK
- 67 interviews were conducted:
  - 20 were from academia and were considered primarily a user of LPS data and had no role in the delivery of an LPS.
  - 23 were involved in the delivery and management of 1 or more LPS study. This included academic study leads, as well as study operational and management leads.
  - 14 interviews with individuals in the delivery of current LPS initiatives that support the delivery or use of LPS.
  - 10 interviews with individuals outside academia, which included representation from industry (3), government (4) and third sector (3).

  Some interviewees spanned multiple roles. Most interviews were with individuals. However, several included small group discussions with individuals from the same organization or study (counted as single interviews)
- Interviews included people in early, mid and senior stages of their careers, and people whose interest is primarily in the social sciences (27) (SS) and health sciences (40) (HS)
- We are grateful to all those that shared their views; these are listed in the appendix from slide 18.

Discovery: Studies

HDRUK
Health Data Research UK

Issues

Numerous interviewees in both health and social sciences cited discovery as a key obstacle in using LPS in the UK, currently absorbing a large amount of time and not always culminating in success. The interviewees cited multiple aspects of discovery, including the ability to

1. Identify the full range of LPS available and the broad scope of each one
2. Understand what datasets are included in each study
3. Understand enough about the context and contents of the datasets to determine whether they are relevant to a particular research question
4. Identify which studies include sufficiently large numbers of individuals meeting particular criteria to be worth accessing in order to answer a particular research question

Regarding the first issue on this list, over 20 interviewees suggested that there was a need for a single discovery site which would provide a discovery function for all LPS across all disciplines. While some spoke of this as a 'catalogue' or 'database', most spoke of it as a 'platform' which would include high quality metadata and a user-friendly search function.

On the other side, some interviewees explicitly suggested that a new discovery platform was not needed, since several platforms already exist, and that the problems with discovery are primarily related to understanding the data contained in LPS.

Potential role for PRUK

The interviewees who thought that a single discovery portal for LPS is required generally felt that this was a role that PRUK should take on. Those who believed that a single discovery portal is not required felt that PRUK should focus on quality of metadata, working with the discovery portals that are already available.

'There are already so many points of access, like UKDS, ONS, SAIL, Closer – we don't need another one!' - SS

'A lot of social scientists don't know what's out there. We had a discussion with [.....] at the ESRC last week and she told us about a whole bunch of linkages that have been created that we didn't even know about.' - SS

'Discoverability is weak. Social scientists don't know where to look for health data, and I suppose it's the same the other way round.' - SS

'Ideally, you wants to be able to go to one place to find data on patients with a set of criteria and find out which cohorts have the most patients with such criteria and then, ideally, receive a merged set of data on such patients from across the cohorts.' - HS

'A single platform would only help marginally – you can usually find what you want from desk research.' - HS

I like the idea of a gateway, but at the moment it still feels just a little bit detached from the people who then give you the data - HS

'We're creating more and more data platforms but the data comes out in the same messy way!' - HS

My major worry is that somebody is doing a research project somewhere and isn't looking at my study – HS

Discovery: Metadata and data

HDRUK
Health Data Research UK

Issues

There was general agreement amongst the interviewees that in many cases it is too difficult and too time consuming to develop an understanding of the data available within a study, of the contextual information necessary to evaluate how the data can be used, and of the key characteristics of the subjects within a study. While some platforms (such as CLOSER) and some studies (such as ALSPAC and Understanding Society) were cited as having made some progress in overcoming these problems, the landscape was seen as very patchy, with the data attributes of many studies remaining opaque to the external researcher.

A common theme was that there are many data dictionaries available, and that additional data dictionaries would not be useful. However, the development of standards for metadata would be helpful – although only if supported by research funders. Additionally, many interviewees pointed out that the development of metadata requires investment, and therefore additional resourcing would be required if high metadata standards are to be achieved.

A solution to developing metadata for legacy data was noted by many interviewees as required but it is a sizeable task and that researchers weren't incentivised to do so. Other interviewees preferred to focus on the opportunity for greeter prospective harmonisation in collection of survey data and metadata.

Several interviewees discussed functionality that would enable a user to assess the data's potential utility (e.g. frequencies and shape of the data) against their research hypotheses prior to data access processes. Interviewees gave examples that could be expanded (e.g. NESSSTAR, DataShield, HDR Innovation Gateway Cohort Discovery and the automatic generation of synthetic datasets).

Potential role for PRUK

Some interviewees felt that PRUK should try to solve the problem of data discovery by providing guidance and setting standards for metadata. This is not a new issue for the field and PRUK should not seek to reinvent the wheel and be interoperable with ongoing activities, A small number of interviewees, with experience of smaller cohort studies, suggested that PRUK should provide either funds or personnel to develop metadata on behalf of studies. PRUK could build a discovery platform but this is not as important as addressing some of the underlying issues around availability and standard of some metadata.

'The ability to understand what's in the dataset varies a lot - Understanding Society is quite good because their website has a search function and filter function. In other cases you have to go to the original questionnaires.' - SS

'You need to present the data in a way that is accessible for social scientists. A lot of them, for instance psychologists, are not very quantitative.' - SS

'National studies have good metadata, but smaller studies have to prepare that metadata. It's not attractive for researchers. It really should be funded separately.' - HS

'I am worried about fragmentation – our data is available through several platforms now and there is a cost to maintaining and updating them all - SS

'What we need is someone to come to us and say, help us understand your data and the best way to present it, and we'll do that for you.' - HS

Documentation of contents of studies is poor, especially for health datasets. - SS

'Some cohorts, like ALSPAC, are good at showing what they have – with others, you have no idea except by looking at what they have published.' - HS

'It requires resource to make data discoverable. It's not very glamorous. Who's going to pay? - HS

Could PRUK have sandpit functionally – so you've got the real interrogation ability to see what the data has in it. - SS

Data access

Issues

Difficulty in obtaining access to data was cited by many interviewees as a significant obstacle in utilizing LPS, frequently taking excessive time because of the duration of the process required to gain access and an excessive burden of documentation. The ensuring delays can substantially reduce the time available for analysis in a fixed-term grant or even lead to analyses being abandoned.

The perception of interviewees was that difficulties in the access process have a number of causes:

1. Under-resourcing within studies and low prioritization of access requests
2. Excessive risk aversion related to the risk of disclosure of sensitive data
3. A 'My data' gatekeeper culture creating an unwillingness to share data
4. Cost of access was mentioned by several users, but less frequently than the time and gatekeeper culture.

It was noted that the degree of difficulty in obtaining access varies substantially amongst studies. It was generally felt that access is more difficult for biomedical studies than studies in the social sciences.

Potential role of PRUK

Some interviewees felt that PRUK should develop access standards or a model access procedure for use across LPS, which would need to be backed by funders to make a difference. These standards might help to achieve mutually recognized accreditation across LPS. Two of the interviewees who thought that PRUK should develop a comprehensive discovery platform for LPS explicitly said that this platform should also facilitate accelerated, streamlined access to the studies. A small number of interviewee thought that PRUK should provide an advisory service to help researchers navigate access processes.

HDRUK
Health Data Research UK

The gatekeepers of the data don't always understand data governance requirements and are overly restrictive because of concerns about patient privacy.' – HS

'The procedures for access are tortuous - almost every grant we have is delayed because of this.' – SS

'People lose applications, sit on them for months, then come back to researchers to ask for further information because they didn't specify properly what was needed at the start!' – HS

'The role of PR UK should be to remove the concern that a project won't get done because the data can't be accessed. This means a single platform with streamlined access.' - HS

' 'Lack of resource often delays the access process. There are so many other things to do when running a study, they just think, access, we'll deal with it later.' – HS

'PRUK could provide standards, guidance and best practice around information assurance and access.' – HS

'It makes the research life cycle very time consuming and demoralising. The balance has gone too far to protecting data.' – HS

'In order to get secure access you need to answer legal and IT questions that a researcher can't answer. It's okay if you can access legal and IT support teams at the university, but students aren't able to do that.' – SS

'It feels like the process is designed to block rather than facilitate access.' – HS

'If you don't know someone from the study, you can pretty much forget it.' – HS

'You just about have to write a paper to get approval.' – HS

'The data that collected in the UK currently is world leading, but the access systems aren't' – SS

'In the end, we gave up.' – SS

Multi-study and cross-study analysis

HDRUK
Health Data Research UK

Issues

The interviewees who have had experience of working with multiple datasets almost all cited access as a particular problem with cross-study analyses (an exception were the UK Data Archive studies e.g. 4 national birth cohorts). The ideal would be a single approval process allowing access to all the required studies, but currently separate approval is generally required for each LPS, creating an additive effect from the problems inherent in seeking approval to access each one.

There was a range of views to harmonization of datasets. Whilst harmonization of variables is resource intensive and it is attractive for this to be done centrally, others commented that they would prefer data in its rawest form possible.

A specific issue when brining together some of the larger studies might be the duplication of participants. It would be useful if there was a way for participants who are in multiple studies to be identified centrally.

Even when access problems are overcome, many interviewees felt that linkage and data harmonisation are under-resourced.

Potential role of PRUK

Since access is such a major problem for cross cohort analyses, the roles suggested by interviewees for PRUK in improving access are applicable in relation to these analyses, especially if PRUK could provide a governance framework that takes account of disclosure avoidance or could facilitate access to multiple cohorts through a single approval process.

'We need to get to a point where we have a trusted framework of ethics approval that is less variable from committee to committee and able to handle multi-cohort consent.' – HS

'My PhD student lost about a year trying to get access to the Millennium Cohort study and in the end they still couldn't do the linkage of data that he needed.' – HS

'There are probably lots of researchers duplicating harmonisation.' - SS

'Getting agreement to do it is glacial. It's probably motivated by fear of data leaks. There's just this massive Institutional intransigence.' – SS

Some people are in more than one study, and PRUK needs a way to help users identify duplicate study participants – HS

I think in social sciences we find the idea joining studies together quite alien, because, again, because what's the population basis that you're doing that on? - SS

Data linkage

HDRUK
Health Data Research UK

Issues

Many interviewees spoke to the times and resources invested in record linkage between LPS and potential health and administrative data sets. Few interviewees were positive about the current status quo, and many expressed frustration at the current challenges in creating linkage.

Where linkages have been established they have not always allowed for onward sharing or have restrictive sharing permissions. Linkages have always been bilateral; multidisciplinary research may want to link across several datasets. The cost of establishing linkages has on occasion has been unfeasible.

Permission for onward sharing is a particular problem with some data controllers. Interviewees felt that in some instances there was also a lack of willingness to allow linkage. Administrative data from each of the 4 nations is not always readily comparable; some suggested PRUK could play a role in harmonizing some key data across national administrative datasets.

Obtaining consent for linked data from participants was not considered a significant challenge, whilst much could be learnt from continuing to share learning and best practice

Potential role of PRUK

Some interviewees felt that PRUK should provide funds for linking datasets, while others felt that PRUK should itself carry out linkage and provide the resulting datasets.

Some interviewees also saw a role for PRUK in advocating for linkage, particularly in relation to government departments.

Researchers in education suggested that PRUK could have a role in developing less disclosive derived variables from linked datasets, which could be made available under a more open licence.

'There's a cultural problem with linkage. It's not really enabled. We have the permissions in place but there isn't the inclination or resource – there's not enough drive to make it happen or it's too siloed and each organisation wants to do the analysis themselves..' – SS

'The Longitudinal Linkage Collaboration will do much of the work PRUK was going to do and enable it to start from a different place.' – HS

'PRUK should become a resource for researchers to access linked data.' - HS

'The problem with linkages is not normally technical, it's usually a super-Kafka problem of one data provider not wanting to provide identifiers or something like that.' – HS

'I was originally told the linkage would take 3 weeks – it's taken 7 months and counting. They're understaffed, but that means I won't be able to do as much as I expected to do before my grant runs out.' - HS

We have been trying to link our study data to health records for 7 years – HS, study rep

Our participants would probably be horrified if they realized how few linkages they'd given consent for had been realized – SS

I'm not sure that practically we can achieve linkages collectively – at the end of the day each study negotiates its own contracts with the data owners – SS

'We're creating these increasingly complex datasets, but we also need to support researchers in using them effectively – SS

We have linkage of health records, but we're not allowed to do onward sharing of that information at the moment, which you know, for us is a really major problem – SS

Providing data and analytical environments

HDRUK
Health Data Research UK

Issues

Data is distributed to users by three main mechanisms (a) by whole download of datasets by a user agreement (UKDS), (b) via the preparation of a bespoke dataset that is sent to the approved user (cohort studies) or for data deemed too sensitive (c) through provision of access in a securely governed safe research environment.

When discussing distribution of data via UKDS, some interviewees did not see barriers to this model being more widely applied. Others highlighted it would not be appropriate due to incompatibility with consent taken from participants, the risk of duplication of analyses, and that without input from the research team it could lead to inappropriate analyses. Interviewees involved in the preparation of datasets for distribution highlighted that this was a resource intensive process.

Interviewees highlighted that accessing secure trustees research environments was an onerous task. Although their necessity was understood, concerns included the time to make access, that the user needed to visit a physical location for secure access, and the availability of analytical tools in some environments. It was noted that the pandemic was accelerating remote access to the ONS SRS.

Several study representatives expressed a concern that the growth of recent initiatives requiring data deposition was potentially leading to fragmentation of their datasets and had resource implications to update and maintain each one.

Potential role of PRUK

For some interviewees, it was seen as crucial that PRUK had a TRE for LPS, and this would enabled streamlined discovery, harmonization, data access and availability of open-source analytical tools and code could be enabled. Two respondents described a model whereby many LPS were deposited and available in the same TRE 'a shared infrastructure'; but that have different 'fronts' offering concierge and services to cater/target different audiences (e.g. dementia, mental health, other topics etc).

Others imagined an expansive network of platforms/ TREs for different types of research including (a) multi-study access (b) linked data (c) biosamples (d) multiomics data and analysis. Some interviewees envisaged PRUK would offer all of these activities, others envisaged elements as a solution to a particular area.

Some interviewees suggested that PRUK could innovate in the generation of synthetic data, which could be available for download to prepare analyses before accessing a TRE environment.

Our data is downloaded 1000s of times via UKDS, so other activities need to be proportional in the resources they need – SS

We need a platform with a number of cohorts, designed for worldwide access, with multimodal data, that's what everyone wants. – HS

We could only now deposit a partial dataset in the (UK Data) Archive and I think that's trouble – HS

We need a small number of big archives that hold data – SS

Other roles suggested for PRUK

HDRUK
Health Data Research UK

Potential role for PRUK

Biosample collections

Several interviewees suggested that PRUK could offer central storage for biosamples. e.g. UKBioCentre – which could offer subsidised storage to studies for their samples. It was also considered this might provide a focal point for engagement with industry with biosample collections.

Omics data and sample assays

Several interviewees suggested PRUK might also fund/enable standardised 'omics (e.g. genomic, proteomics, metabolomics) across many LPS collections. However, there was considerable challenge in harmonising omics data across studies, due to variability arising from different sample collection approaches.

Legacy studies

Some interviewees also referred to 'mothballed' LPS and wondered if PRUK might seek to play a role in ensuring that the data in these studies remain available for use. Several study representatives specifically discussed the challenges of maintaining biosample collections as high expense after a study closed.

Sharing resources

A number of interviewees spoke of the need to store and share code and other tools that are developed to analyse, derive or harmonise data so as to reduce duplication of effort and enable analyses to be undertaken more quickly. They wanted to see PRUK create a repository for such tools. Where studies lack these or other tools (examples given included code to enable researchers to combine sweeps in a cohort study and an e-consent tool), PRUK could develop these rather than each study developing the tools for themselves.

Collaboration

Several interviewees also wanted to see PRUK encourage collaboration and sharing of experience between studies – one gave an example of creating a central resource for survey questions. Many interviewees said that they valued the opportunities which CLOSER provided in convening the community and providing opportunities to learn from each other.

We're known as a social science study, but we do have genetics, epigenetics and the like – it would be helpful to raise awareness of this –SS

We have a large sample collection to maintain, which needs about £10k per year, but its incredibly hard to raise these funds – HS

Individual studies should not be negotiating to convert their samples to data – HS

Biobank has been Gamechanging in showing how data and samples can be provided as a service – HS

PRUK could support in making sure that we're that our methods are harmonized - HS

Other roles suggested for PRUK

HDRUK
Health Data Research UK

Potential roles for PRUK

Tracking use of data

Several LPS study representatives spoke about the challenges of tracking impact from LPS. The challenges were **different** across health and social care. A lack of consistent dataset identifiers for health datasets was mentioned; by contrast there was not a project register for ongoing analyses in social sciences.

Training

A number of interviewees thought that PRUK should be involved in providing training to researchers and data scientists in the use of LPS or in encouraging more training to be incorporated in post-graduate courses. Areas of training included data handling; e.g. how to prepare and transform data for analysis and using complex analysis and datasets (e.g. approaches to working with linked data). Others suggested this could be a point of differentiation between PRUK (platforms and infrastructure) and CLOSER (training).

User support

Others spoke to the needs for more practical user support to help users with navigating the use of LPS. E.g. helping users navigate the LPS and their metadata, design analysis and supporting access processes.

Interface with policy makers

Several interviewees spoke about working at the interface with government and policy makers. For government, analysts often work on rotation and expertise moves around, policy questions have short turnarounds. The prospect of deeper concierge services to analysts was attractive; or tools to rapidly visualise data by ethnicity or demographics.

Public advocacy

A small number of interviewees suggested that PRUK should play a role in advocating for LPS with the public. The interviewees that mentioned this felt it was very high priority.

Could PRUK help map our use and user footprint for when we have to report or make a submission for new funding - HS

There's there just seems to be this unending appetite for training data. If you're a platform that's sharing data, particularly complicated data would have thought training would be a good thing - SS

PRUK could have people who could look at the sorts of research questions government departments have, and alert them to data sources of which they might previously been unaware.

Coming off the back of the greatest ever, modern exposure of the general population to epidemiology, communication to the general population about the utility of being part of studies and allowing your data into research would be an absolutely brilliant thing to achieve - HS

Other points

Funding

- A number of interviewees mentioned the fact that funding for many LPS, particularly in the biomedical sciences, is short term, while the studies themselves are envisaged as continuing over decades. They felt that a different, more strategic funding model for these studies is required, and wondered if PRUK could play a role in achieving this
- Several respondents highlighted that PRUK would need to be supported by a bigger drive from funders for multidisciplinary research. E.g. focused calls that incentivised novel and broad use of LPS.
- Several interviewees pointed out there is a continued need for research funders to support the individual studies. PRUK is a 'meta-activity' that relies on investment in fundamental data collection

Scientific approaches between social and biomedical sciences

- Several interviewees noted that there were fundamental differences between the studies run by social sciences and more clinically led studies; for instance, the importance of the sampling framework chosen by studies. Also, many biomedical studies do not ask rich social science questions and vice versa. E.g. a researcher interested in family dynamics might find limited data collected in some studies that they could use.
- The pooling of studies was seen to have different value between health and social sciences; several interviewees spoke about the power of bringing studies together for subgroup analyses whereas individual social sciences studies are adequately powered for the majority of topics of interest.

Working with other infrastructure and initiatives

- A number of interviewees, excluding those that envisaged PRUK providing a single discovery platform for all LPS, wondered how PRUK planned to work with other organizations in the field such as UK Data Archive, CLOSER and DPUK

Appendix 1: PRUK design and dialogue development programme interviewees

Name	Organisation
David Newton	Akrivia Health
David Crosby, Anbalakan Paramasivam, Talisia Quallo & Alexis Webb	Cancer Research UK
Michael Dale	Department for Education
Rachel Connor, Leanne Dew, John Wilkinson, Lorna Langdon, Beth Bradley	Department of Health & Social Care, Science, Research and Evidence Directorate
Mike Daley	Department work and pensions
Thomas Keane	EMBL-EBI
Emma Gordon	ESRC
Adam Steventon	Health Foundation
Michael Jones	Institute of Cancer Research
Manish Patel	Jiva.ai
Paul Ashley	Johnson & Johnston Innovation
Seeromanie Harding	King's College London
Guy Goodwin	NatCen
David van Heel	Queen Mary's University of London
Frank Kee & Angie Scott	Queen's University Belfast

Name	Organisation
Charlotte Neville	Queen's University Belfast
Nigel Slone	Sollis
Yvonne Kelly	UCL
Oliver Duke-Williams	UCL
Dorina Cadar	UCL
Andrew Steptoe	UCL
John Jerrim	UCL
Lisa Calderwood	UCL
George Ploubidis	UCL
Spiros Denaxas	UCL
Emla Fitzsimons	UCL
Aida Sanchez	UCL
Rebecca Hardy	UCL
Natalie Banner	Understanding patient data
Nic Timpson	University of Bristol

Interviewees

Name	Organisation
Deborah Lawlor	University of Bristol
Jonathan Sterne	University of Bristol
Joanne Newbury	University of Bristol
Andrew Boyd	University of Bristol
Carol Brayne	University of Cambridge
John Danesh	University of Cambridge
Adam Butterworth	University of Cambridge
Rob French	University of Cardiff
David Porteous	University of Edinburgh
Robin Flaig	University of Edinburgh
Cathie Sudlow	University of Edinburgh
Michaela Benzeval	University of Essex
Gundi Knies	University of Essex
Matthew Woollard	University of Essex
Lee Elliot Major	University of Exeter

Name	Organisation
Jason Gill	University of Glasgow
Annette Boaz	University of Kingston
Peter Diggle	University of Lancaster
Janet Cade	University of Leeds
Martin Tobin	University of Leicester
Olly Butters	University of Liverpool
Richard Kingston	University of Manchester
Prof James Banks	University of Manchester
Fiona Matthews	University of Newcastle
Sheena Ramsay	University of Newcastle
Paul Burton	University of Newcastle
Daniel Stow	University of Newcastle
Phil Quinlan	University of Nottingham
John Gallacher	University of Oxford
Frank Windmeijer	University of Oxford

Interviewees

Name	Organisation
Hazel Inskip	University of Southampton
Colin McCowan	University of St Andrews
Ronan Lyons	University of Swansea
Dave Ford & Chris Orton	University of Swansea
Madeleine Murdoch	University of Glasgow

Online survey outputs

Summary: online survey

- An online survey was launched between 5 March and 9 April 2021 as part of the PRUK design and dialogue development programme. The survey was developed in Survey Monkey and asked respondents about their current experiences of working with longitudinal population studies and their priorities and ideas for the role of Population Research UK.

- This was a convenience sample. It was promoted twice to a mailing list of 252 people whom had signed up to receive updates on Population Research UK following the PRUK development launch webinar. It was also promoted in Hive – a monthly newsletter of HDR UK, and via social media of HDR UK, ESRC, MRC and Wellcome.

- The survey had a maximum of 27 questions which included multiple choice, rating scale, and open questions. 216 respondents responded to the questionnaire. 107 completed the questionnaire in full, and 109 provided some partial response. Respondents were predominantly UK academic audience that was broad across geography, numbers of years experience in working with of Longitudinal Population Studies (LPS) and disciplinary background.

Summary: Current experience of using LPS data

Summary findings

- Respondents were asked about their current experiences of using and working with LPS. Experiences of using LPS vary widely across participants. The areas of most common challenge with working with multiple studies and linked datasets. Respondents also highlighted access to appropriate training.
- LPS Discovery is variable across datasets and disciplines. Currently challenges with understanding contents of datasets and getting access to data can cause studies to be delayed or abandoned.
- A quarter of respondents indicated that that would plan to wait 6 months or more to get access to data.
- 82% participants would like to work more with LPS linked to a broad range of health and administrative datasets (82%, 111/136). 40% of respondents cited difficult access as the primary reason, 22% a lack of awareness of data available, 10% skills development, and 6% data quality. Other reasons included cost of accessing data.

Summary: Ideas and priorities for the role of PRUK

- Respondents highlighted a wide number of functions that PRUK might take on. The most commonly identified and highest priority functions across the survey were:
  - facilitating data linkage to a variety of health, administrative and other data types
  - streamlining data access
  - making it easier to use datasets in combination
- Respondents highlighted a number of opportunities for PRUK to tackle long-standing issues challenges in longitudinal population studies, such as the representativeness of current studies, incentives for data sharing, the different approaches and policies of research funders, perceived lack of investment in studies and the need for analytical methods skills and training to support complex data analysis.
- The survey outputs support the direction of travel of PRUK with a focus on streamlining data access and data linkage. In developing PRUK, respondents highlighted the need for PRUK to work with and build on good practice. There is a need for initiatives to be more joined up and interoperable. The survey highlighted that at a national level there must be the right balance between investment in LPS to collect, manage and share data and meta-initiatives such as PRUK.

Methods: sample characterisation

- Respondents were almost entirely based from the UK (97%; 200/208), and predominantly worked in academia (~81%; 168/208). There were a small number of respondents from government and public sector (4%; 9/208), third sector (4%; 9/208) research funders (4%; 8/208) and members of the public (4%; 8/208).
- There was a good mixture of academic disciplines captured by the respondents; 96 respondents worked in epidemiology or public health, 71 people identified working in the biomedical or health sector and 47 from economic and social sciences. (number greater than 168 since respondents could indicate 2 responses).
- 144 provided their place of work, which covered different 36 universities and 12 further employers.
- Respondents had a range of prior experience working with longitudinal population research. Around a third of the respondents indicated they had been involved with LPS for >5 years (33% 61/184); 50 people (27%) had involvement for 5-10 years and 73 (40%) had been involved for more than 10 years.
- Respondents were involved in the generation, use and management of LPS data. A quarter of participants had experience in the management or delivery of an LPS (26%, 51/196). Two thirds (66%; 131/196) of respondents had experience of analysing LPS data, and around a fifth (22%; 44/196) involved in initiatives that supported others in the use of LPS. (respondents could select multiple ways to be involved)

Current experience of using LPS data: overall

Key points

- Respondents were asked to indicate how they would rate their experience (poor, fair, average, good, excellent) with different steps of using UK LPS data.
- Respondents reports variable experiences across all stages of using LPS data.
- Respondents identified complex data access requests (multiple data sets or requests for linked data sets) as the area that there was most challenge. It was noted by some respondents in their remarks that there was more available linked health data compared to other types.
- Experiences of discovery and access on this single question were rate more positively than suggested elsewhere through the qualitative interviews. This suggests that there are examples of good practice to build on.
- The completeness and consistency of data available was rated as one of the areas where participants had better experiences.

Data type	% rated poor, fair	% rated good or excellent
Discovery	18	48
Data access	22	45
Accessing multiple datasets	36	18
Accessing linked health data	37	18
Accessing linked administrative data	42	10
Completeness and consistency of data	20	48
Access to appropriate training	24	29
Dissemination	14	45

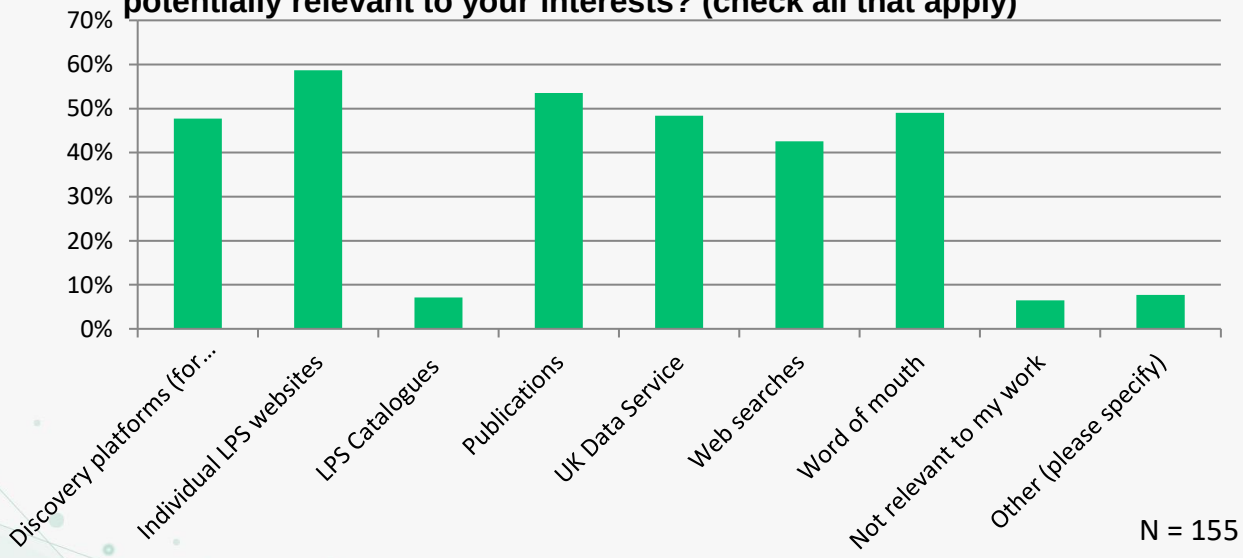
Q8: N = 148 (68 skipped question); see annex 1 for graph

Current experience of using LPS data: discovery

Key points

- ~48% of respondents had indicated a good or excellent experience with discovery (Q6).
- Respondents use a broad range of approaches to discover LPS datasets (Q8). This may reflect personal preference, or that there isn't a catalogue with listing of all LPS.
- Respondents did highlight challenges with understanding the contents of datasets. The frequently had caused delays and abandonment of planned analyses (annex 1)

Q8: What methods do you use to discover LPS datasets that are potentially relevant to your interests? (check all that apply)



N = 155 (61 skipped question)

Respondent comments

CLOSER has done a great deal of work improving discoverability. Government/ public sector respondent

Difficult to find information in one place. Academic (multiple disciplines)

Access through UKDS usually good documentation. Though sometimes need to download data in order to understand variables and sample sizes – Academic (social & economic sciences)

I find the data dictionary useful - but only after 2 years of working in the area to understand what metadata might be needed. Academic multiple disciplines)

Improved since the HDR Gateway. Academic (epidemiology)

Many good resources like CLOSER but more work could be done to create compatible data dictionaries and code repositories. Academic (biomedical and health sciences)

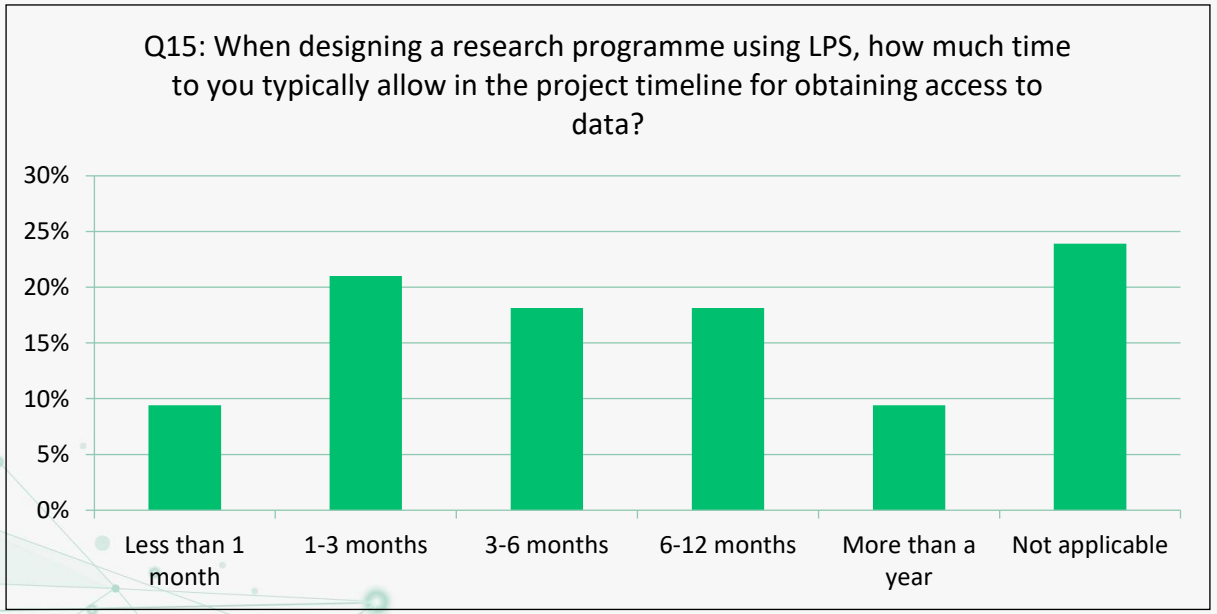
I have discovered that there is commonly data available to some but not all, or data that is discoverable if you know where to look. Academic (social and economic sciences)

What is lacking is a central place where you can see what data might be available from what studies that can help with your research. Academic (epidemiology)

Current experience of using LPS data: access

Key points

- Experiences of data access are variable – specialist data services that provide use were reasons cited for more positive experiences
- On the other hand, challenges with data access were have found to often cause delays and abandonment of analyses (data in annex 1)
- Whilst some responses were critical, others highlighted that processes although lengthy and subtly different studies were generally well described.
- These challenges were compounded when making multiple study requests



N = 138 (78 skipped question)

Respondent comments

Every study has different access rules which are not transparent. Academic – (epidemiology, study delivery)

*Great support from Celsius officers. A*cademic – (social and economic sciences)

I associate the difference with whether the data is held by social or medical scientists - Academic (health)

Each data access is well documented and at an individual level, simple to access. Academic – LPS study delivery and management

Having access to data via services such as UK Data Service makes use of data much easier and more convenient – Third sector

Although the process of accessing data is lengthy it is generally clear, and support is sometimes offered by the data controller themselves - Academic – (epidemiology)

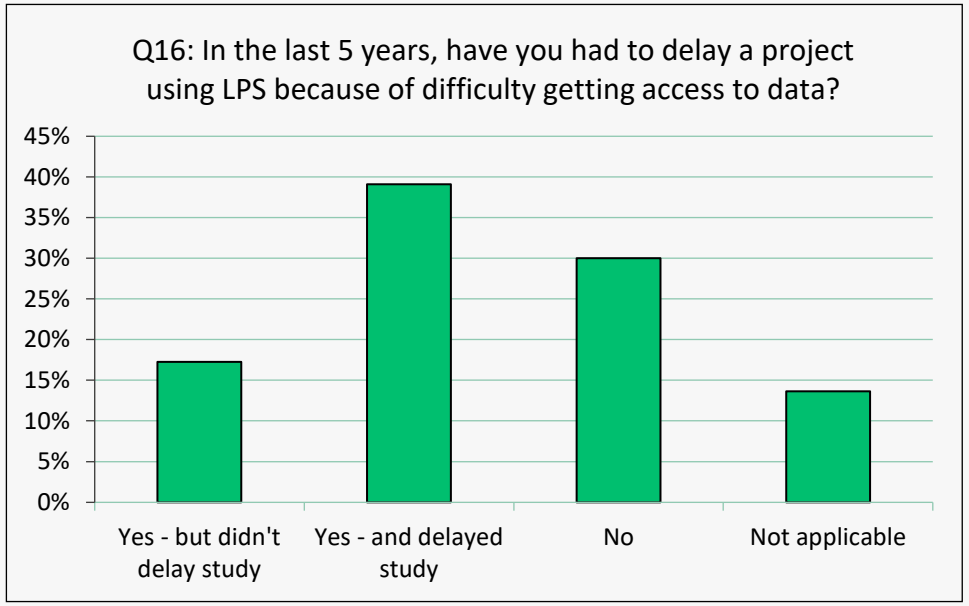
Further support, for example through a network of researchers/mentors who previously used the data, would benefit new researchers in particular - Academic – (epidemiology)

Data access is OK, but requires funds to be able to request data. Academic (health)

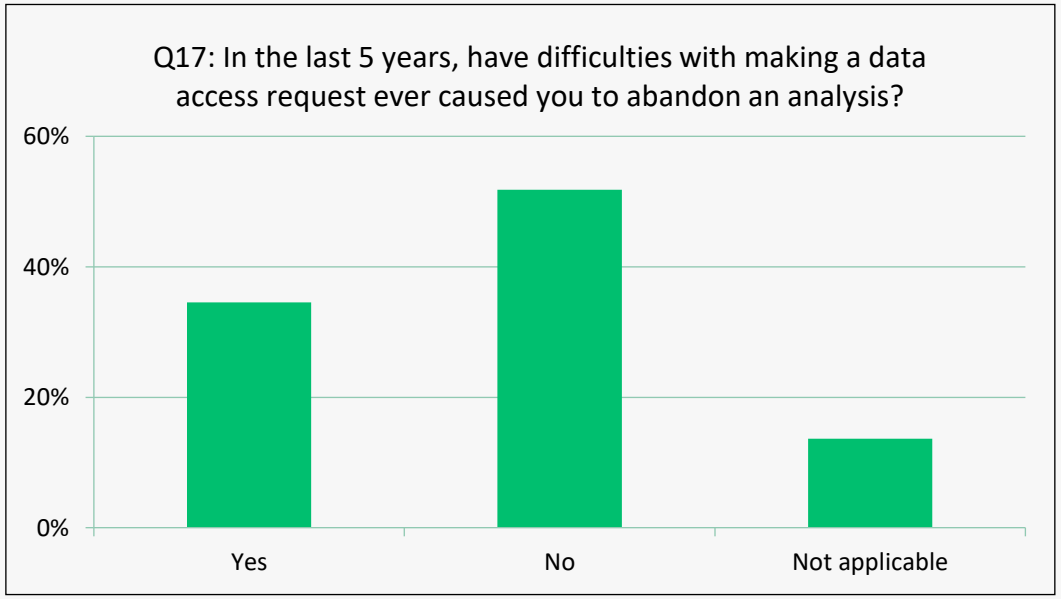
There is a cumbersome process that is unique to each study. The UKDS for social science studies suggests the benefits of a TRE approach. Academic (study delivery)

Difficulty of accessing multiple datasets due to confidentiality issues. Academic (epidemiology)

Data Access:



N = 110



N = 110

Current experience of using LPS: linked data

Key points

- Accessing linked health or administrative datasets was rated the lowest experience by respondents (Q6). Only ~10% of respondents reported a good or excellent experience in linking to administrative data and 20% to health data)
- There is a strong appetite for working more with linked datasets (82%, 111/136) (source Q19 (n=136). 40% of respondents cited difficult access as the primary reason, 22% a lack of awareness of data available, 10% skills development, and 6% data quality. Other reasons mentioned included no current need and cost. (source Q20 (n=106).
- Respondents would like to link to a broad range of data types across health, social and economic and environmental data. (see Q21). The most often mentioned was health data, but financial, educational and environmental data was mentioned by a number of respondents

Q21: What data would you particularly value being linked to LPS? (where n >5 responses)

Data type	No of responses
Health (general)	27
Financial/ economic	13
Secondary care	12
Social care	12
Primary care	11
Education	9
Geospatial (various)	7
Environmental (exposures)	6
Mortality	5
Wearables	5
Local authority	5
Administrative (general)	5

Q21 N= 73 (free

Respondent remarks

I really value working with linked data, but I will never do it again as it took 3 years

I would like to access existing health data and link with all types of environmental data and this is difficult, expensive and often impossible. Academic (epidemiology)

(rated excellent). We work with ADRS and linkable admin data in Scotland, e.g. education. Academic (biomedical and health)

Heavy regulatory burden to access e.g. linked ONS data. Academic (biomedical and health)

barriers from data owners mean limited linked data available Academic (social and economic science)

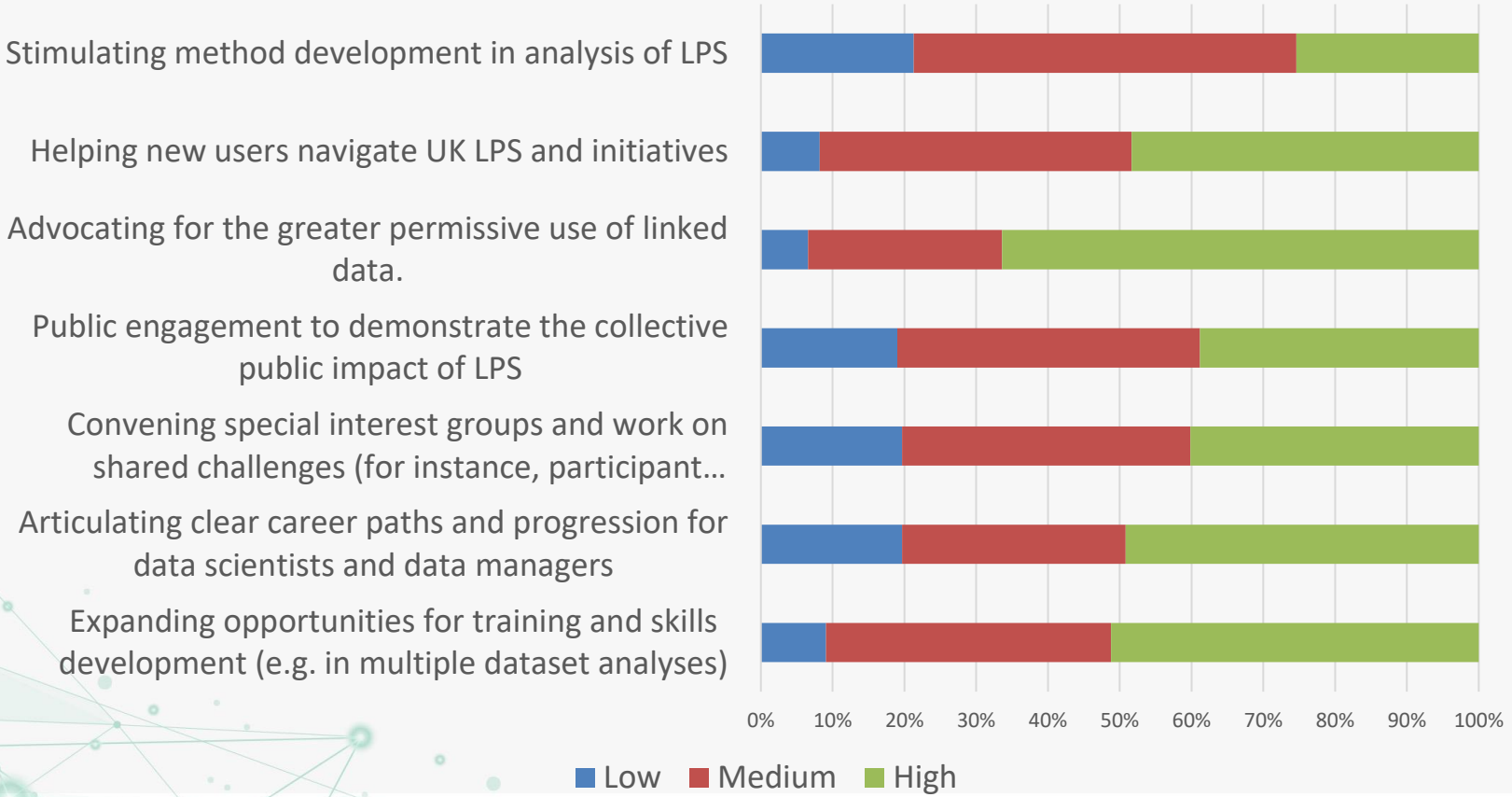
People I advise have seldom had any training at all in the governance aspects of using LPS, especially in relation to linkage Academic (health)

Priorities and ideas for PRUK

Key points

- Respondents identified a range of high priority roles that PRUK could play. PRUKs role in advocating for use of linked data was rated highest priority by the most respondents.

Q22: Please assign a priority to each of the following activities for where you think PRUK should prioritise its activity for supporting collaboration between researchers, studies and organisations engaged in LPS



Respondent remarks - what else might PRUK prioritise

- *Clear career paths and progression are essential for all -what about technicians, laboratory managers, field workers and general managers?*
- *Availability of smaller grants focused on enhancing LPS*
- *Allow descriptive stats to be generated (e.g. variable distributions) without the need for extensive data access approval*
- *Collaboration to increase quality of social care and social work data*
- *There should be cross-cohort working groups around the introduction and standardisation/harmonisation of new phenotypes (e.g. wearables, administrative data, genomic/proteomic/metabolomic data, etc.)*

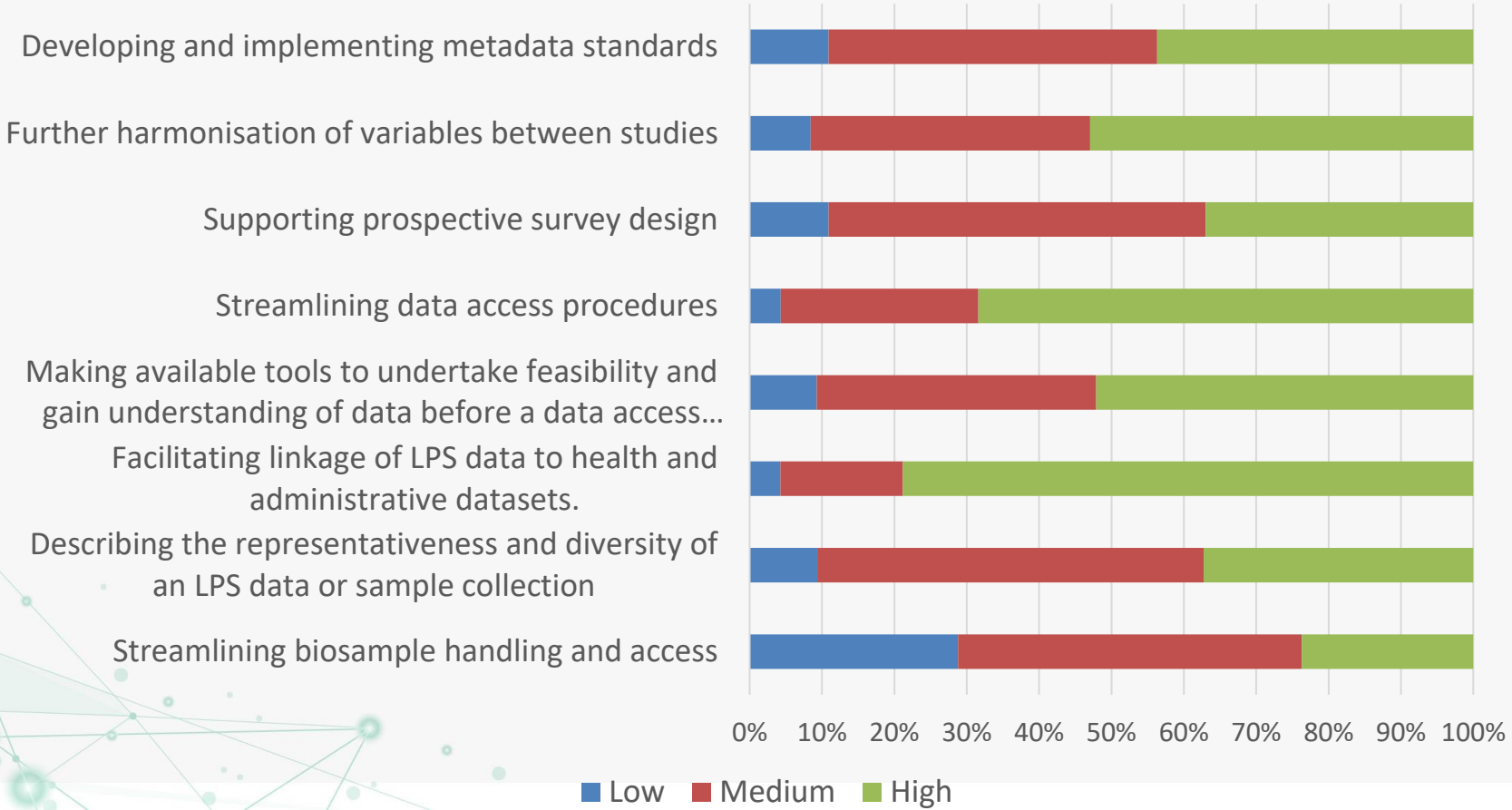
N = 122 (94 skipped question)

Priorities and ideas for PRUK

Key points

- There are many functions regarded as high priority that PRUK could develop. Respondents consider streamlining access (70%) and facilitating data linkage (80%) as highest priority activities for PRUK.

Q23: Please assign a priority to each of the following activities for where you think PRUK should prioritise its activity for increasing the use of LPS data and samples.



Respondent remarks – what else might PRUK prioritise

- *Having dedicated funded technical support who could support researchers in navigating the complex procedures*
- *Really important that existing efforts to champion communications of studies and the value of LPS is done.*
- *Use of open source Analytics & data visualisation tools & training*
- *Documentation of methods used in the LPS, and their consequences for analysis*

N = 119, 97 skipped question

Priorities and ideas for PRUK

Q25: Please describe what you see as the greatest single opportunity for increasing the contribution of LPS to accelerating impact for research and innovation across the UK (freetext).

Theme	Responses	Respondent remarks
Facilitating data linkage	9	• *Reduction in bureaucracy for linkage to e.g. NHS-Digital or PHE data* • *Linking LPS to other data sets*
Streamlining data access and governance	9	• *Help researchers navigating through the different datasets and help in getting access to data* • *Facilitating access and visibility to multiple, multidisciplinary researchers, with data sharing and collaborative work prominent*
Bring data together	6	• *Having a joined up programme of LPSs which enables questions to be answered across datasets* • *Enabling datasets to be used in combination*
Attention to method and data collection	5	• *Greater reproducibility and robustness for population science* • *Ambitious datasets usually mean poorer quality samples and such complex designs that people don't analyse them appropriately.*
Funding new LPS / new datasets	5	• *Multigenerational family studies* • *Generating multidimensional data from molecular to intricate phenotyping*
Improved descriptions and awareness of data	4	• *Better description of what the data is really like before spending months applying for it.* • *Making sure researchers know about what data is available*
Increased funding to longitudinal studies	3	• *Provide resources to the LPS themselves to be able to engage fully* • *Improving the underlying issues (e.g. underfunding, incentive issues)*
Greater public involvement and engagement	3	• *Engaging the public in the value of LPS to get more support for research* • *Greater public presence and involvement of public panels in publicity working*
Focus on specific topic areas	4	• *Prevention of disease, and factors influencing trajectories* • *figuring out how to build back better post-covid*
Focus on translation	3	• *Link it to policy - e.g. areas of research interest (ARIs)* • *Faster translation into practice*
Other	1	• *A coordinated approach to analysis of biological samples* • *training the next generation of researchers* • *Work with industry; Learn from Biobank*
Not classified	4	• *What is LPS again?*

Priorities and ideas for PRUK

Key points

In free text remarks, a number of respondents highlighted some overarching challenges in longitudinal population research that PRUK could seek to address. These included:

- Representation of UK population in current national LPS portfolio.
- Variability in processes between studies and lack of agreed standards (e.g. metadata) creating lack of interoperability between data, studies and initiatives.
- Differences in requirements between funders in policy and funding models.
- Legacy data issues. E.g. resources and investment required to bring up to standards that are useful for stakeholders.
- Different costs and data access decision processes.
- Lack of funding or incentives for some data sharing and management activities in some areas.

Several respondents highlighted that the survey was overly focused on the use and linkage of data and they were not asked questions to give a perspective for those of us who work on designing, maintaining and supporting the original studies. Key to the success of PRUK will be developing a valued and meaningful partnership with longitudinal population studies

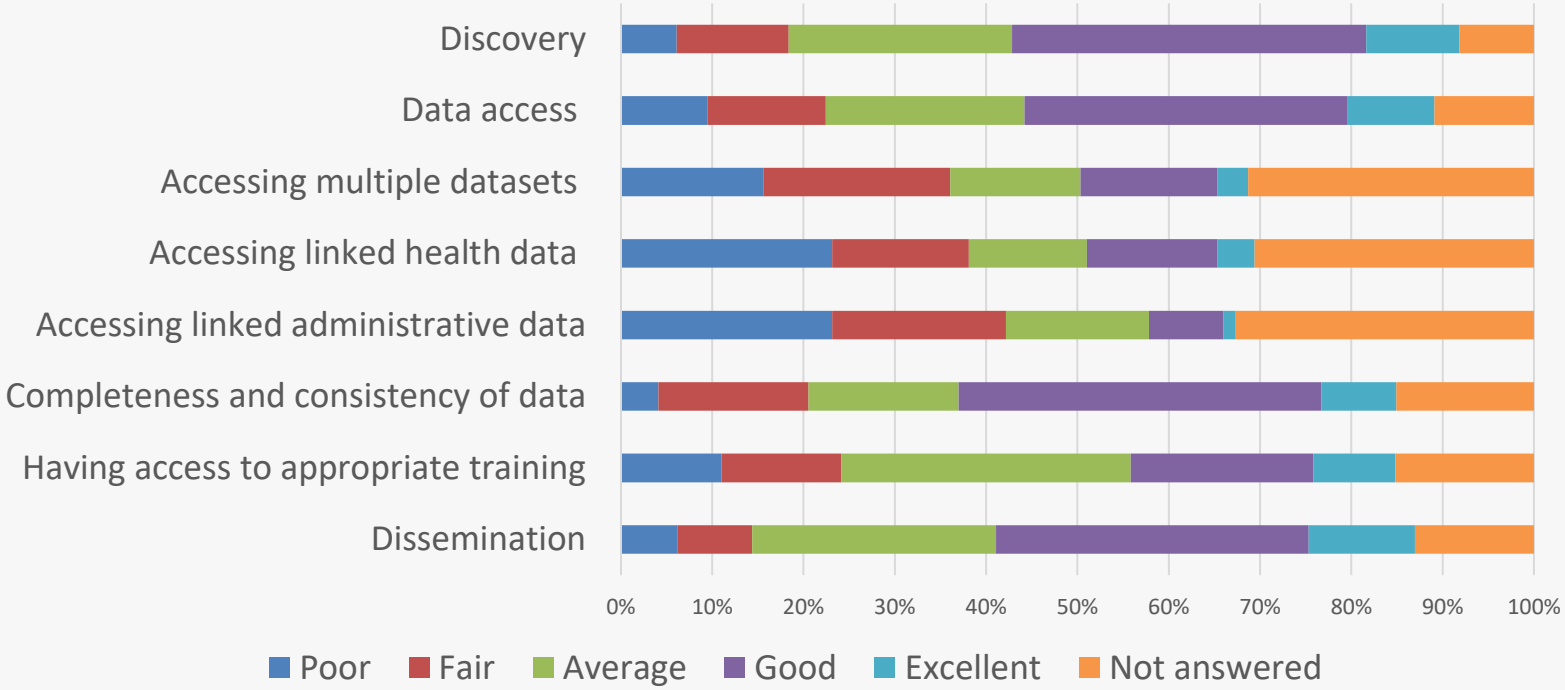
Respondent remarks

- *Investment in specialist teams (through the retention of expertise and training of new staff) to make study data accessible and discoverable across funders is essential.*
- *The problem also goes beyond just investment from the different research councils and there are cultural differences in access arrangements and expectations around discoverability of data.*
- *There is a need for a metadata strategy across the PRUK.*
- *LPS should be properly funded, not least to optimise and fix historic data*
- *There is a large incentive problem in contributing to the scientific development of LPS. They are under funded and contributing to them properly takes a lot of time.*

- *PRUK should build on, and support, existing expertise and best practice (especially within ESRC-funded infrastructures e.g. UKDS and CLOSER) to retain an international reputation.*
- *PRUK should create a roadmap for interoperability within and between disciplinary domains, whilst encouraging further innovation.*
- *Any new resource/infrastructure must be interoperable with existing resources.*
- *Access and discoverability of ESRC-funded LPS is already at a high standard through resources like UKDS, CLOSER Discovery and some study websites. MRC/Wellcome funded LPS do not have the same history of investment in data access and discoverability, so are facing a very different problem.*

Annex 1

Q6: Please indicate how you would rate your experience (poor, fair, average, good, excellent) with each of the different steps of using UK LPS data.



N = 148 (68 skipped question)