Longitudinal Population Studies

Data Audit Report 2022

Jon Johnson (Technical Lead)
Alice Barnett (Research Assistant)

8 November 2022

Acknowledgements

I would like to thank:

- The study PIs, data managers and others who participated in the survey.
- Alice Barnett who ably assisted with the setting up of the survey, data management and help with the analysis; and
- The PRUK Development Group for guidance and feedback on the report.

Executive Summary

The purpose of the data audit was to provide a baseline of information to enable the ESRC and MRC to understand the current state of the many activities that the Longitudinal Population Studies (LPS) carry out, to better understand where gaps exist, where improvements may be possible and to plan for future data collection and support for UK LPS. Sixty-nine studies were identified to be invited to complete a survey, these included 64 longitudinal studies, and 5 cross-sectional (see Appendix A). Fifty-one (74%) of studies responded to the survey. Analysis of the response by primary funder (ESRC or MRC) and by size of study indicates that those that responded are sufficiently representative of UK LPS as whole, to draw robust and meaningful conclusions.

From 2017-2021, over 111,000 dataset requests were received by the 51 studies; these include bespoke datasets and standard datasets such as those available from the UK Data Service (UKDS). The number of datasets being shared with researchers by studies varies widely. Small studies averaged around 100 datasets, medium sized studies around 1,100, large studies around 3,600 and cross-sectional studies around 15,000 datasets.

There has been a marked rise (36%) in the number of datasets being accessed by researchers over this five-year period. Though this has varied across the different studies, this pattern holds irrespective of the size of the study, or where the data is shared from. Use of third party Trusted Research Environments (TREs) such as the UKDS Secure Lab and SeRP platforms expanded 2-3-fold. The number of dataset requests declined across the board during 2020, with the exception of those from the UKDS End User Licence (EUL) infrastructure; this is likely due to the hiatus in data access in TREs in the early stages of the pandemic. Numbers began to return to their pre-pandemic levels in 2021, and there is no reason to expect that the overall upward trend is likely to diminish, especially for data held in third-party TREs.

Alongside this, the number of infrastructures the studies are expected to share data to is a concern. GDPR-related withdrawals could additionally increase the frequency of updates, and studies expressed a strong preference for providing data to a single data provider, who would then be the ‘canonical source’ for data at other infrastructures.

The main barriers for linkage of LPS data with data from other sources were “obtaining consent from the data holder” followed by “implementing governance arrangements” and “delays in data provision” once linkages were available. There were also concerns about the timeliness of data availability to researchers, in part related to TRE capacity to carry out manual output checking. Development of a shared output disclosure strategy and processes would assist in reducing these barriers. The audit findings indicates a near doubling of planned linkages to health and government department sourced administrative data; this points to a need to put in place better mechanisms for studies to interact with departments, to reduce the lag and burden of data linkage application and delivery. There is valuable work being carried out by the studies assessing the quality of administrative data by comparison with LPS data.

Established best practice such as the common safe researcher training program and accreditation aligned between Office for National Statistics (ONS), UKDS and HMRC could provide a firm basis for a scheme that would extend to data held in other TREs.

Data sharing of large datasets, in particular omics data, scans and other images, is almost always done using the studies’ or its institutions' resources. In the absence of a shared secure infrastructure with appropriate governance and sufficient storage and compute, it is difficult to envisage a viable alternative location for these linkages in the short to medium term. This poses barriers not only to cross study analysis, but it limits the possible analysis due to limitations in the resources available at the study. Small studies are particularly affected in this regard.

UK LPS hold significant and valuable samples, but currently many are not registered in the UK Clinical Research Collaboration (UKCRC) catalogue; adding them would raise the visibility and encourage more usage. Governance and sample depletion policies across studies vary widely, however there is much good practice which could be used as the basis for a consistent set of information which is made publicly available to researchers.

There were in total 16,700 publications reported between 2017 and 2021 from the 46 studies which have collected data, illustrating the important role longitudinal population studies play in supporting research across a wide range of scientific domains. A number of studies have changed their policy regarding authorship, and arrangements for data access to researchers outside of the immediate research group, for instance by re-consenting participants. Sixteen studies (31% of those who responded to the survey), reported that co-authorship was still a requirement. We did not investigate whether any of these studies could be in a position to have a more inclusive policy this would be worthy of further follow-up.

Individual studies have created a number of different discovery solutions to provide researchers with information to guide researchers on what data are available and provenance information (e.g. questionnaires, data collection description). The variation in the content and mode of delivery reflects the available skills, the infrastructure at the study or its host institution and the number of users of the data. This presents a fractured discovery experience for researchers coming to LPS investments. Cross study resources such as CLOSER Discovery[1], UKDS[2] or Gateway to Global Ageing[3] provide a more consistent experience but lack coverage across the full range of studies. Ninety five percent of all data shared (from 2017-2021) by LPS investments is discoverable from three metadata infrastructures with complete metadata, CLOSER Discovery, Gateway to Global Ageing or UK Biobank[4], in total these infrastructures represent 32% of LPS participating in the survey. Investigating ways such as federated discovery across these infrastructures could be beneficial, but there should also be investment to make the scientifically valuable long tail of studies made more discoverable.

In the short term, there is a need for a single point of entry for these investments, which can provide basic catalogue information, to guide potential users to the 'right place' along the lines of an enhanced MRC cohort directory[5]. In the longer term, a metadata strategy for infrastructures and studies could start to address ways in which best practice could be developed to encourage more interoperability in both technical standards and interoperable content between studies to meet the differing needs of those looking to use LPS data.

As previously noted, studies were open to modernising data management infrastructure and processes to support better discovery and data production. There is a wide range of skills, software and personnel across the studies, and there is scope for sharing best practice even within the existing resources. The lack

[1] https://discovery.closer.ac.uk
[2] https://ukdataservice.ac.uk
[3] https://g2aging.org/
[4] https://www.ukbiobank.ac.uk/
[5] https://www.ukri.org/councils/mrc/facilities-and-resources/find-an-mrc-facility-or-resource/cohort-directory/

of resources and good guidance has made it difficult for some studies to move away from legacy systems to support more modern software and hardware stack. Most studies reported recruitment and retention issues and as such are vulnerable to loss of (knowledgeable) key staff.

Recommendations

Data Sharing

1. Need for coordination and alignment of access and governance policy with current and planned data holders outside of health and ONS to meet the needs of LPS, to inform the technical and governance architecture of TREs.
2. Development of a shared output disclosure strategy and processes.
3. Promote the contribution of LPS to assessing the quality of administrative data.
4. Studies should be supported to improve and modernise data management processes and infrastructures including incentives for using shared resources where possible.
5. ESRC and MRC should investigate the development of guidance on best practice for data access and governance to ensure a consistent set of information is made publicly available.
6. Establish a network or community of practice to share experience, develop skills and best practice would be welcomed by many, but especially by the smaller studies, to take forward any guidance.
7. Investigate mechanisms for providing data to a single data provider, who would then be the ‘canonical source’ for data at other infrastructures.
8. The audit indicates a near doubling of planned linkages to health and government department sourced administrative data; this points to a need to put in place better mechanisms for studies to interact with departments to reduce the lag and burden of data linkage application and delivery.
9. The common safe researcher training program and accreditation aligned between ONS, UKDS and HMRC could provide a firm basis for a scheme that would extend to data held in other TREs.

Samples

10. Encourage studies to register samples at UKCRC to enhance the discoverability of samples in UK longitudinal studies.
11. ESRC and MRC should investigate the development of guidance on best practice for sample governance and sample depletion policy to ensure a consistent set of information is made publicly available.

Metadata

12. Establish a metadata office to develop and oversee a metadata strategy which sets out clear guidance for studies and data dissemination infrastructures on minimum content and metadata format.
13. In the short term, there is a need for a single point of entry for these investments, which can provide basic catalogue information, to guide potential users to the ‘right place’ along the lines of an enhanced MRC cohort directory. In the longer term, a metadata strategy could start to address ways in which best practice could be developed to encourage more interoperability between studies to meet the differing needs of those looking to use LPS data. Topics used in such a directory should be revised to better reflect the diversity of studies.
14. A high-level catalogue should be developed and regularly updated and maintained as a central resource for high level information about the coverage of topics, data and study profile of UK LPS studies.
15. Investigate the possibility for federated discovery and interoperability across existing infrastructures.

Findings

The UK has a long tradition of collecting longitudinal data, spanning participants born over the last century. This covers studies commencing before the advent of readily available computerisation through to the modern day, and a huge diversity of populations, study designs and scientific rationale. Studies are also diverse in the level and stability of funding; their capacity to provide best practice in areas such as data management and respond to the changing expectations for data sharing. The purpose of this data audit was to provide a baseline of information to enable the funders to understand the current state of the many activities the studies carry out, in order to better understand where gaps exist, where improvements may be possible and to plan for future data collection and support for UK Longitudinal Population Studies.

Sixty-nine studies were identified to be invited to a survey, these included 64 longitudinal studies, and 5 cross-sectional. Fifty-one (74%) of studies responded to the survey. Analysis of the response by primary funder (ESRC or MRC) and by size of study indicates that those that responded are sufficiently representative of UK LPS as whole to draw robust and meaningful conclusions. Five studies were in start-up or a piloting phase. All other responding studies held participant data from surveys, 96% had some form of data linkage and 83% held samples from participants.

Terminology

Through this report, the following terms are used. Infrastructure is used to mean an investment by the funder, this could be a single study, a platform such as the UK Data Archive, or a Trusted Research Environment (TRE). Where a specific type of infrastructure is relevant, this is stated explicitly. There are multiple TRE environments based on the Secure eResearch Platform (UKSeRP), these include the Secure Anonymised Information Linkage (SAIL Databank), Dementias Platform UK (DPUK) and the UK Longitudinal Linkage Consortium (UK-LLC); for tabulation purposes, these are often categorised into SeRP Platforms.

Data sharing

The number of infrastructures through which participant data is available to researchers is likely to increase, currently 45% share through more than one infrastructure, and 21% through more than two infrastructures, the primary driver for this is to link to other data sources.
Studies primarily funded by ESRC all share data through the UK Data Service, though many additionally share data in other ways, for example to TREs, or directly as bespoke datasets. Studies funded by MRC primarily share data directly to researchers as bespoke datasets. Table 2 in the Descriptive Report illustrates the different combinations of data infrastructures studies are sharing data through.

Those studies which are predominantly sharing data directly to researchers are overwhelmingly doing so as bespoke datasets. The primary drivers for producing bespoke datasets are data minimisation and disclosure control. Studies also reported that providing an advice and guidance service to identify the most appropriate variables was highly valued by many researchers as part of the bespoke dataset creation process.

Studies reported the resource intensive nature of producing bespoke datasets on both researchers and the study data managers. These studies are characterised as having small numbers of participants and sharing either as downloads to the researchers with Data Transfer Agreements (DTA), or increasingly through local, normally institutionally run TREs.

From the risk perspective, placing data in an institutional TRE is seen as a positive development and the direction of travel; however, the data management resource needed is barely reduced and the overhead of on boarding data into a TRE is only marginally, if at all, less resource intensive than issuing a DTA. In follow up interviews with some of the studies with these characteristics, there was a desire to modernise and improve data management processes and infrastructure, but the resource was already a fractional post with little room for additional development time. These studies also indicated that they are not well placed if there was a significant increase in the number of data requests.

Studies reported on the challenges of sharing data through multiple infrastructures. The main concerns were versioning of data. Many studies refresh research data on a periodic basis, the increased awareness of GDPR amongst participants has necessitated removal of cases and subsequent reissuing of datasets, some studies re-release the entire dataset to incorporate a new data collection, and there are also occasions where data is reissued for corrections to longitudinal variables, household demographics etc. Studies expressed a strong preference for providing data to a single data provider, who would then be the 'canonical source' for data at other infrastructures. This, however, can be a complicated process, as it could require sharing of identifiers between infrastructures. An additional consideration is updating of any accompanying metadata, as there is little interoperable metadata, necessitating reissue of metadata in different formats.

Over the period 2017-2021, there were over 111,000 datasets requested or downloaded by researchers from the 51 studies; these include bespoke datasets and standard datasets such as those available from the UKDS (which may be downloaded several times by the same researcher) The British Election Study accounted for 60,000 data requests (54% of the total). Table A shows the number of data requests handled by the different infrastructures and the number of studies using each. Because some studies use multiple infrastructures to share data, the total adds up to more than the total number of studies.

Although the number of data set requests is not strictly comparable across all infrastructures (UKDS distributes data for many studies as multiple datasets - normally one or more per data collection event, whilst others provide a single bespoke dataset) it provides an insight into the profile of where data is being shared from by the studies.

Table A. Data sharing by infrastructure

Infrastructure	% data requests	No. of studies
Directly from Study	65.7%	41
- British Election Study	53.7%	1
- UK Biobank	9.8%	1
- Other	4.0%	39
UKDS (EUL/SL)	25.9%	15
UKDS (Secure Lab)	0.4%	9
SeRP platforms (DPUK,SAIL,UK LLC)	3.1%	11
Other	3.1%	8

Of the studies under consideration, there has been a marked rise (36%) in the number of datasets being accessed over the five-year period. Though this has varied across the different infrastructures, this pattern is also reflected when looking at the number of requests by study size. Use of third party TREs such as the UKDS Secure Lab and SeRP platforms has expanded 2-3-fold. The number of dataset requests was down across the board during 2020, with the exception of those from the UKDS End User Licence (EUL) infrastructure - this is likely due to the hiatus in being able to provide data through TREs at the early stages of the pandemic - but numbers began to return to their pre-pandemic levels in 2021, and there is no reason to expect that the overall trend upwards is likely to diminish, especially for data held in third party TREs.

Although the numbers start from a small base, there has been a sharp increase in the numbers of datasets being accessed.

Time to access data shows a wide disparity between studies. Immediate access is available only from studies using the UKDS (under End User Licence) and the British Election Study which employs a similar licence to the UKDS. Nine studies are able to provide data within a month, and 10 within 3 months. Three studies reported that it takes between 3 and 6 months and 2 studies more than 6 months. Fourteen studies were not able to provide a timescale, on the basis that it varied depending upon the complexity of the enquiry. Medium sized studies were more likely to share data in a timelier manner, although the majority of this was accounted for by those using the UKDS.

Data is made available to researchers in a variety of formats. 60% provide data in SPSS or Stata. These formats by definition provide additional information on the data, such as variable descriptions, format definitions and, where appropriate, labels for codes. Other formats such as CSV/TSV (50%) or other mentioned such as SQL or ASCII would need accompanying documentation.

We asked studies about barriers to data sharing currently and those they anticipate. The number of studies anticipating funding of infrastructure and staffing as becoming a barrier in data sharing is respectively 6-fold, 3-fold higher than the number currently experiencing them as a barrier. Rate of study growth and challenges in sourcing staff, training and/or expertise were also identified as anticipated barriers.

We asked studies about barriers to data management. Staff funding is markedly the most common concern amongst studies, ~50% of studies have identified it as a barrier to data management both currently and in the future. Other barriers reported include staff recruitment reported by ~30% of studies along with staff training/expertise and data sharing funding both reported by ~20% of studies. These studies do not anticipate these barriers changing in the future.

Omics Data & Images

Seventy percent of studies have omics data available. A wide range of omics data is being shared. As with participant data, the overwhelming number of studies shared omics data directly with users. Many though use additional infrastructures for some data types. There are 30 studies which are sharing omics data solely through their own resources. Figure 12 in the Descriptive Report illustrates more detail on which other infrastructures are being used.

The information has limitations as we did not ask about the volumes of omics data being shared. Data shared through the European Genome-Phenome Archive (EGPA) will be exclusively omics data; much of the data shared directly from studies is also likely to be of this nature. Sharing omics data linked to phenotype data will normally be shared directly from the study as the data holders will be the custodians of the link between these data and the only organisation able to enact this linkage.

Image data is available from 50% of studies. A wide range of different types are made available, as with omics data, shared directly from the study. DEXA and MRI scans are the most widely held image type, where image data is held within the NHS it can be shared with that infrastructure, but in the majority of cases they are held at the study.

Data Linkages

Data linkage to longitudinal population studies not only enhances the possible types of analysis of the study participants, but it is also able to act as a benchmark for the quality and robustness of administrative data.

The largest number of linkages are to health and spatial records, followed by education and employment. Studies also link to external agencies such as NHS Digital and DfE to have up-to-date address information for participant tracing. Table B shows the number of planned and possible linkages. If these were to be achieved, it would nearly double the number of linkages to the studies.

Table B shows the likely upper bound of linkages for the studies. Not all linkages will be possible, records may not exist for the time period, there may not be consent or indeed the area of linkage may not be adding value to the study, e.g. consumer behaviour to a study focused on genetics.

Table B: Data Linkages Overview

	Current		Planned		Possible	
Linkage	No of studies	Linkages	No of studies	Linkages	No of studies	Linkages
Health	26	47	12	24	10	16
Education	8	13	10	25	5	9
Employment / Income	4	8	2	6	11	27
Criminality	2	2	3	3	3	6
Spatial	20	49	6	17	2	3
Consumer	3	4	1	0	0	0
Other	10	12	1	1	2	6
Participant tracing	9	9	0	0	2	2

As noted previously, studies have an optimistic view about the prospect of health linkages, or at least to NHS Digital data; there are a large number of planned applications and approvals and potentially almost as many in the more distant future. This will have resource implications for both the studies and NHS Digital itself. A similar statement could be made about education linkages and although more distant, there are many possible linkages to data from the Department of Work and Pensions, and HMRC.

Table C provides a high-level overview of which infrastructures data linkages are available from. As with participant data, there are a large number of infrastructures being used, sometimes for the same data linkages.

Table C: Current data linkages by infrastructure

Data access infrastructure	Health	Education	Employment	Criminality	Spatial	Other
Directly from Study	24	9	12	1	19	11
UKDS	6	5	0	0	3	3
SeRP Platforms (inc UKLLC)	9	2	4	1	2	1
NHS/NHS Digital	11	0	0	0	0	0
Other	21	11	0	1	7	4

These reflect the likely increase in both the number of data linkages and of the diversity of infrastructures where these would likely need to be supported.

The biggest challenge to emerge from the data audit for data linkages and more specifically any data held in a TRE, is that of scaling data sharing. The capacity for TREs to ingest and link data has made significant strides over the last 10 years, aided by the building of relationships between the data holders and the providers of secure infrastructure.

The phase one report from DARE[6] provides a number of recommendations for a roadmap for the technical architecture for TREs to support trustworthiness of these infrastructures to lay the basis for federated data and easing of movement of data between them, primarily in the health domain, but also including ONS. This data audit has identified that there is also a need for an alignment of governance policy by a wider range of data holders including government departments and agencies to support the current and envisaged linkages to longitudinal population studies, for the technical architectures such as those developed under the DARE program to be operationalised, and to inform and engage with the different concerns of the various data holders.

The increase in planned linkages, and the secular trend to place more data in TREs, poses a significant challenge to the management and governance of data access, and output checking for disclosure. There is a common safe researcher training program and accreditation aligned between ONS, UKDS and HMRC; this would provide a firm basis for a scheme that would extend to data held in other TREs.

However, the capacity to significantly increase the number of researchers access to a TRE to conduct analysis is primarily constrained by the ability of the TREs to provide sufficient resources for output checking. This is a complex problem from a technical perspective as there is a wide range of software used for analysis resulting in different formats of output to be checked, and thus not amenable to automation.

[6] https://dareuk.org.uk/wp-content/uploads/2022/08/DARE_UK-Paving_the_way_coordinated_national_infrastructure_sensitive_data_research-Aug2022.pdf

In the short term, TREs are in a challenging environment for the recruitment, training and retention of staff to carry out manual output checking

There is a proliferation of infrastructures, which pose a number of governance, technical and resourcing challenges for studies. We asked studies to rank the major barriers for data linkage. "Obtaining consent from the data holder" is overwhelmingly the most onerous. Followed by "implementing governance arrangements" and "delays in data provision" from the data holder.

For studies which are conducting data linkages directly with data holders, this poses significant overhead for studies in gaining approval for the linkages, managing the data holder governance arrangements. This is more keenly felt by smaller studies with less internal capacity, especially if dealing with multiple data holders with different arrangements. Once agreement is in place, funding is required to carry out the data linkage and extraction. Studies reported difficulties in determining the timescales for delivery of data from many data holders, in some cases the period of agreement expired and the whole process had to restart. From the studies' perspective, this makes it difficult to schedule a window of opportunity to carry out the data preparation work needed on receipt of linked data from a data holder, and managing expectations to curate the data in a timely manner for use by researchers, with little ability to plan effectively.

Where data linkages are held or conducted in third-party infrastructures, data holder policies mandating that linked data is held in a nominated infrastructure mean that there is no effective mechanism to allow equitable movement between TREs.

An illustration of this is the complications of conducting analysis across country borders. English education linked data (for new linkages) from DfE is allowed to be held only in ONS-SRS, Welsh education data at SAIL, Scottish education data at UKDS Secure Lab. A similar situation would arise for instance where participants in a Scottish study moved to England or Wales whose NHS data would be available in UK LLC would not be available in a Scottish TRE alongside the majority of data from a Scottish study. These scenarios are not limited to cross-national boundaries, a similar situation would arise where linked data from different data holders was held in UKDS Secure and ONS-SRS.

For studies primarily funded through ESRC, researchers have a long and trusted relationship with the UKDS for the provision, discovery and use of EUL and secure data. Whilst there is common accreditation between UKDS and ONS-SRS, there is not a reciprocal relationship for the movement of data between them.

An unintended consequence for studies which have a cost recovery model is that providing data to a third party which is free at the point of use, although good from a researcher perspective, means that the cost recovery fees (which are used to backfill data management costs, retain staff and expertise) are no longer accruing to the study.

Samples

The UKCRC Tissue Directory[7] was established to help researchers discover samples and data, help resources improve their data systems for sharing, and harmonise policy relating to the discovery and use of samples and data. Of the studies which held samples, 22% were registered. Two thirds of studies had a depletable resource management policy.

[7] https://biobankinguk.org

The most common sample types were blood (31 studies), saliva (18 studies) and urine (15 studies).

Studies with blood and saliva samples were, in absolute terms, less likely to have registered with the UKCRC.

Follow-up interviews and analysis of the narrative responses indicate that many studies could be more aware of the services that the UKCRC Tissue Directory provides, especially in those cases where the samples were held by a collaborator or a third-party tissue bank.

The 2015 EAGDA report, Governance of Data Access[8] recommended that: "funders should set expectations that studies will develop clear policies on the management of depletable resources, ensuring guidance and support is provided to study leaders in this process".

Although a third of studies with samples did not have a written depletable resource management policy, a review of those with policies indicated that they were written in a way that provided broad principles for accessing the applications. For example, "All applications to use samples should demonstrate a clear scientific rationale regarding why the study is appropriate to the proposed research, and for non-renewable samples, that the use of samples is justified by the expected contribution to the scientific body of knowledge". Follow-up interviews indicate that those studies without a policy did have guiding principles, along similar lines, but these were not publicly available.

We could find only one study, Biobank, with an equivocal policy on coverage "the assay should be conducted on all 500,000 participants or, at the very least, from a large sub-set of randomly-selected participants", albeit with an exception clause for specific use cases.

There is very good guidance on best practice for collecting, storing and managing samples[9]. There is little on what constitutes best practice for a good governance policy and what areas it should explicitly contain, for instance the studies' position on case control vs case cohort approaches, coverage, assay, depletion and output criteria.

Discoverability and metadata

Studies have created a number of different solutions to provide researchers with information to guide them on what data is available and provenance information (questionnaires, data collection description etc). The variation in the content and mode of delivery reflects the available skills, the infrastructure at the study or its host institution, and the number of users of the data. Studies with smaller numbers of data users tend to have less sophisticated provision of metadata than studies with large numbers of users.

The coverage of metadata provided by studies were for datasets (69%), variables (47%), individual questions (46%), summary statistics (24%) and keywords or vocabulary (27%). Questionnaires were mostly provided as PDFs (63%). Table D gives an overview of the format in which these different types of metadata are made available.

[8] https://eprints.whiterose.ac.uk/92286/

[9] https://www.ukri.org/about-us/mrc/our-policies-and-standards/ethics/

Table D: Provision of metadata across all types

Metadata format	HTML/webpage	PDF	Excel	API	Other
% of studies	43%	33%	18%	10%	6%

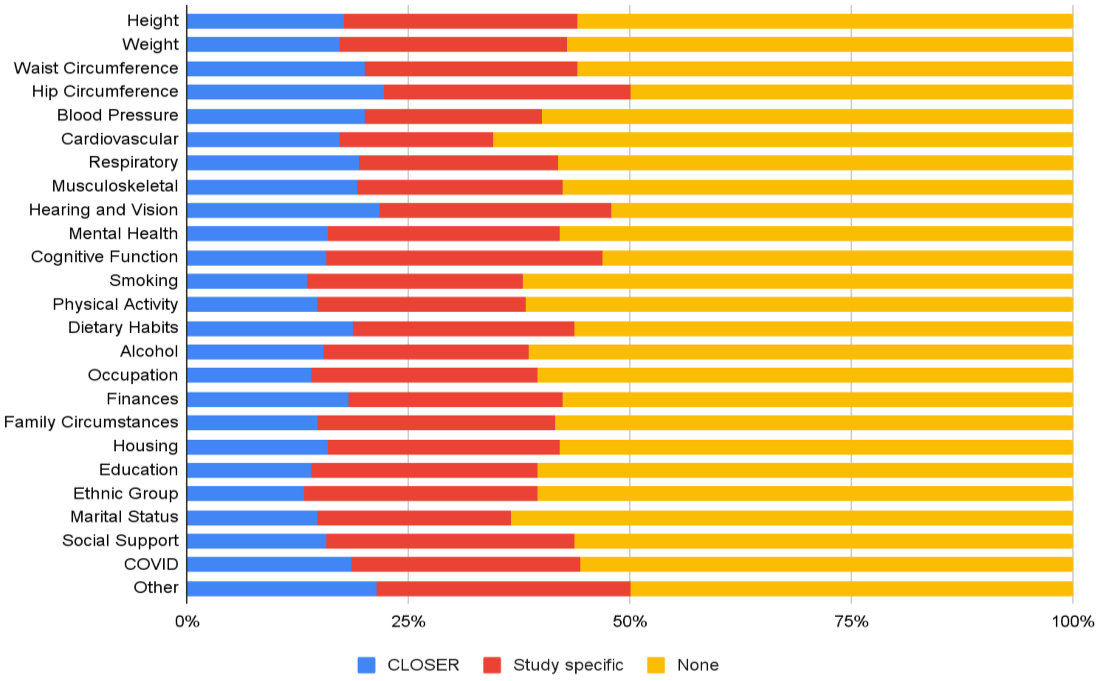
* Figures add up to more than 100% as studies provide information in multiple formats

The categories chosen, HTML, PDF, Excel, API are indicative of the accessibility for discovery of these metadata resources. Provision of metadata in, for instance, Excel once it is located, is very useful for tasks such as filtering or subsetting to select variables of interest. But a PDF or Excel document will have a different utility to a searchable web page (especially if it is indexed).

The studies have a wide range of topical coverage. The topical categories used in this survey were drawn from the MRC Cohort Directory and show the commonality of topical areas which indicates both the breadth of coverage of these studies, but also the potential overlap for cross study comparison.

Figure A combines the audit survey questions on topical coverage with those that reported having keyword metadata and splits that by those which are in CLOSER Discovery and as such use the same list of terms, and those which use other lists of terms.

Figure A: Topical coverage by vocabulary



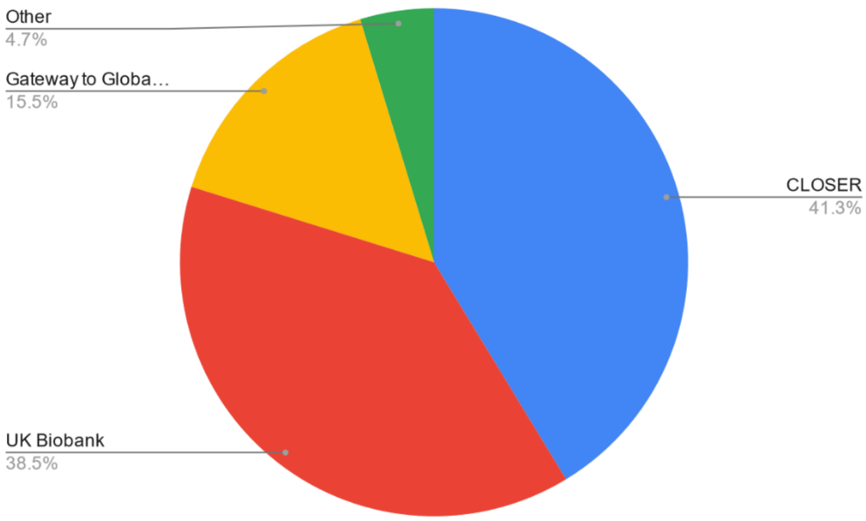
Tagging variables with a consistent topic can be a significant driver for discovery of data; searching just the text in variable labels or indeed questions if they are available is suboptimal for finding data. Even high-level terms will give users confidence that they are finding all the available data for a particular subject. Fifty-one percent of studies hold no variable level topic information, so there is no consistent way in which users can discover the topics associated with the data being shared, i.e. there is no topic available with any metadata which might be available. Twenty-two percent have a study specific vocabulary and 15% utilise the CLOSER vocabulary

The purpose of improving discoverability is to increase the visibility of available data to as wide a possible audience and to inform potential users about the suitability of the data for their research question. There is however a broad relationship between the scope and size of the study and its potential number of users. It would be unreasonable to expect a study with a few hundred participants which addresses a very specific question to have as many data users as a very large and/or multipurpose study. The corollary of that is the level of investment which can be justified in proportion to the potential number of data users.

Figure B shows the number of datasets shared which are discoverable for studies with a full set of publicly available metadata, including summary statistics (e.g. frequencies and number of cases) and keywords. These are only available on three infrastructures, CLOSER Discovery, Biobank and Gateway to Global Ageing. In total they enabled discovery of 95% of LPS data over the years 2017-2021.

The studies included in other, may not be in a position to publish complete metadata, due to concerns about disclosure as they are in most cases smaller studies, although there may be scope for publishing a reduced set of summary statistics.

Figure B: Discovery by data usage



It is helpful to consider the user journey for discovery. The first step could be considered data availability; once that has been established there would be an assessment of data utility, or data exploration and then a decision about data access. These different steps in the user discovery journey require different levels of metadata content.

From a study perspective, journal articles remain one of the primary routes by which researchers discover the availability of data, for instance the use of cohort profiles. For studies with low volumes of users there is little incentive to invest in discovery of data beyond this, and as such they are happy with managing user enquiries about the most appropriate data on an adhoc basis, providing data dictionaries and further support e.g. using data buddies to develop data access applications.

However, for studies with larger volumes of data users, there is considerable advantage for a researcher, as a first step, in being able to locate data for the topic of interest, and adoption of an existing vocabulary would be a beneficial initial step and would lay the basis for integration into a future cross-study catalogue.

The user journey for data exploration is currently fragmented. Studies such as Biobank and those in CLOSER have public access to full summary statistics to evaluate the distribution of the data at a variable level, some studies also provide this information, after a registration process. But for the majority of studies, no information is provided on the number of valid cases at a variable level, which makes assessment of the data's suitability difficult to ascertain. There may be concerns that such information may be disclosive especially for small studies. Provision of minimal information such as the number of valid cases would assist researchers prior to sometimes what may be a burdensome data application.

The discovery landscape is very confusing for potential users of UK LPS data. Multiple infrastructures hold partial information on what data is available in a range of different metadata formats and with varying levels of content. The slimming down of the MRC Cohort Directory so that it is no longer searchable has removed the only (albeit not actively updated) place, with an overview of a good range of LPS data.

The development of a metadata strategy, which sets minimum standards reflecting the researcher user journey for what should be available for each study, would be a first step to creating a cross-study catalogue. New infrastructures would have a set of expectations on the content and the metadata format they would provide, and studies would be in a position to develop a data management infrastructure and processes which could ensure that that information was provided in a sustainable manner.

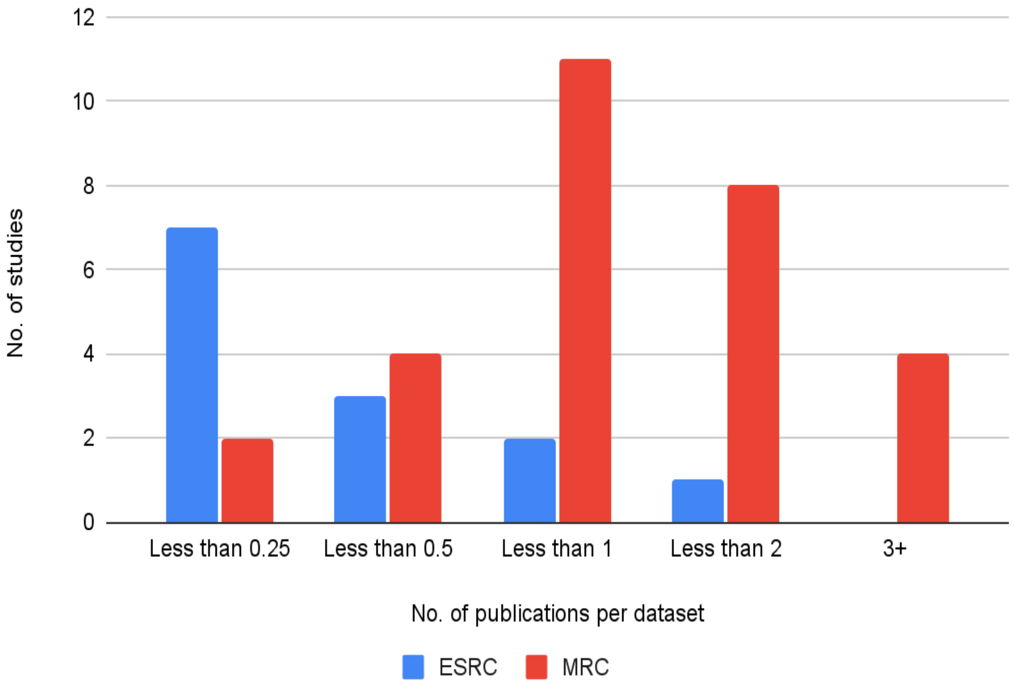
Publications

There were in total 16,700 publications reported between 2017 and 2021 from the 46 studies which have collected data. Making direct comparison between studies at the aggregate level is probably not very helpful, some are small studies focusing on very specific subject areas, whilst others are both more general "all purpose" studies. And some studies have recently started, and time from data request to publication varies between subject areas and journals.

We asked studies to split the publications by whether they were co-authored with the study team, or solely with external authorship. Publications using small studies were far more likely to be co-authored. Half of the total number of publications by small studies is accounted for by three studies.

Figure C illustrates a "crude" measure of whether dataset downloads lead to publications, over the five-year period. The figure shows the split between primary funders, as almost all ESRC funded studies provide data through the UKDS. There is seemingly a higher "rate of return" in terms of publication for each dataset shared for MRC funded studies. What the figure most likely represents is a difference in behaviour of researchers using the UKDS vs requesting bespoke datasets. Researchers are likely downloading data for data evaluation purposes and as noted previously the total number of datasets being downloaded from UKDS will inflate the denominator.

Figure C: Publications per dataset (2017-2021)


No. of studies
Less than 0.25
Less than 0.5
Less than 1
Less than 2
No. of publications per dataset
ESRC
MRC