UK Longitudinal Linkage Collaboration
Population Health Sciences
Bristol Medical School
Oakfield House
Oakfield Grove
Bristol
BS8 2BN

The Longitudinal Linkage Collaboration: a platform for longitudinal COVID-19 research.

Research Protocol.

Public

Version 1

27th October 2020

1. Executive Summary

UK Government and the Chief Scientific Advisor are establishing a funded programme of National Core Studies for COVID-19 research including the Longitudinal Health & Wellbeing National Core Study (LH&W NCS). LH&W NCS will establish a centralised, responsive research resource for the investigation into high priority COVID-19 (C-19) research questions. We will combine a diverse range of high value data assets including established longitudinal population studies (LPS), clinical studies and C-19 volunteer studies from across the UK, in addition to whole-population health and social records.

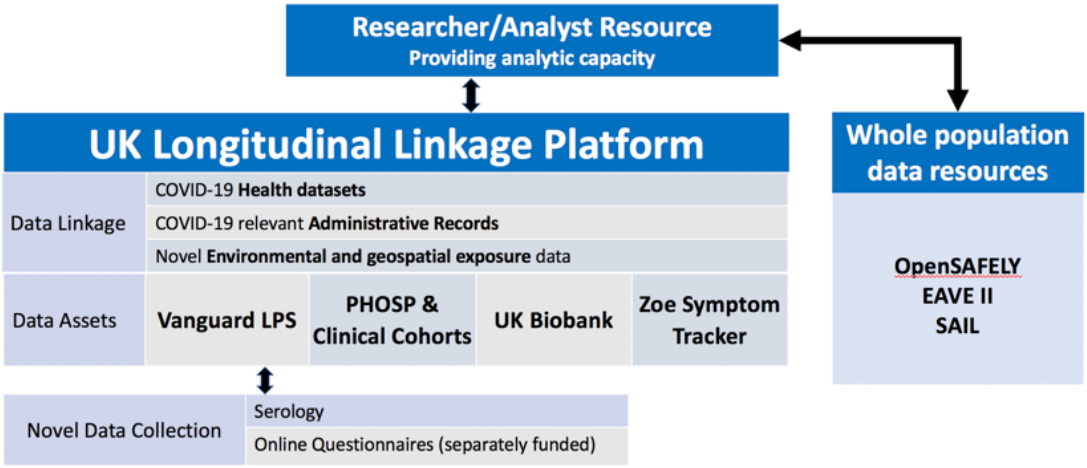
Our vision is to underpin the NCS programme by taking the unprecedented step of combining data assets from 15 major inter-disciplinary UK LPS and utilising these alongside the UK Biobank, the Zoe Symptom Tracker cohort and potentially other relevant sources (Figure 1).

LPS sources of data include biological sample collections, genomic data and in-depth and self-reported measures of health and wellbeing collected before and after C-19. The UK Longitudinal Linkage Collaboration (UK LLC) will develop a centralised research resource where LPS in the UK LLC can be co-analysed. By combining the studies in this way including data from volunteers of all ages, social backgrounds and ethnicities and providing significant geographical coverage of the UK. Within the UK LLC the LPS data will be enhanced by linking self-reported data to volunteers' routine health and administrative records (big data sources). This model allows the investigation of associations identified in whole population records through deep-dive analysis of rich LPS data. Through this the UK LLC will establish a longitudinal linkage platform to host these data assets within a secure environment and link in novel data collection (administered through contributing studies), geospatial exposure data and health and administrative records within a secure environment

These data assets will include detailed measures of C-19 related outcomes and pre-C-19 baseline datasets. The UK LLC will be a unique resource that aligns established data infrastructures with a diverse range of ground-breaking, UK research studies to inform C-19 research and policy.
The UK LLC will be updated from data assets that will continue to collect and record new data, including biosamples, providing insight into the rapidly changing health and social consequences of C-19. Furthermore, combining efforts in this way is much more cost-effective than developing new frameworks.

We will undertake the data integration and management processes needed to align these varied data to support consistent, transparent and defendable research analysis. Aligned with this, national datasets will be used to assess coverage, bias and representation within the sum population of the contributing studies. We will do this through leveraging infrastructure and governance components established by the Health Data Research UK (HDRUK) programme and Administrative Data Research UK (ADRUK)/Office for National Statistics (ONS), Open Safely and the extensive coordination and involvement of the LPS community.

Figure 1: The UK LLC platform for longitudinal C-19 research



To develop the UK LLC and to deliver the LH&W NCS research programme, this study will involve collaboration with the Immunology and Clinical Trials NCS to provide long-term follow-up of participant outcomes. The UK LLC platform is designed to fit within the HDRUK Alliance and will utilise outputs of the Data & Connectivity (D&C) strand which will create/enhance key infrastructure elements ('Trusted Research Environments', national C-19 datasets, data discovery and access tools).

2. Lay Summary

The UK has a long-standing tradition of conducting 'longitudinal research' in order to inform decisions about health care and social policy making. Longitudinal studies work by selecting a group of individuals or properties and then repeatedly collecting data on these people (or the people living in the properties). The groups of people – known as 'cohorts' – are typically selected by having something in common, such as cohorts of pregnant women or people living in a certain area or of a certain age range. The value of longitudinal research lies in collecting a broad range of data, and then repeating this data collection at regular intervals. This allows researchers to investigate the interactions between different things that occur in individuals lives and how changes in this can lead to changing health status or personal circumstances. It is estimated that over 3 million people in the UK take part in a longitudinal study, and many take part across a lifetime – the National Study of Health and Development is still collecting data from babies born in one week of 1946.

The study data are held within research 'databanks'. Researchers wanting to investigate a particular research question have to apply to access data from these databanks. The researcher will need to demonstrate how their research will improve the public good and how they will maintain appropriate standards and information security. For many years the data provided by study participants has been enhanced by 'linking' data collected by the study to data in participants health and other routine government records (such as school records). This provides new information which is hard to collect directly from people and data which can improve the accuracy of study findings. It is also a way in which groups who find active study involvement difficult can be included in the research and therefore benefit from the value research brings.

This project is developing a new approach for linking longitudinal studies to routine records and to provide a secure computing setting where these data can be used in research. We will establish a centralised mechanism for linking records which will be run by the University of Bristol and the University of Edinburgh. This will mean the data from each contributing study can be analysed with data from other studies. By doing this we will further increase the accuracy of the research and enable researchers to study small population groups in more detail. All the data being used will be first processed by the original study so that participants identifiers (such as name, NHS ID) are removed. The data will be stored and analysed in a specially designed secure research computer. No data can leave this system and researchers using the system are audited to make sure they are following the rules. There are governance processes in place to make sure that the data can only be used to help improve the public good, the data cannot be used for profit making purposes. All the original studies will continue to control the data from their participants and will have a role in the research approval process.

This work is being conducted now in order to inform the UK's COVID-19 research programme. Longitudinal studies are able to help in a unique way. They are collecting new data during COVID-19. This can help identify and possibly explain patterns in COVID-19 status and also the effects of lockdown on physical and mental health. It can also help understand the interaction between health and wellbeing relating to the home, work and personal finances, and the neighbourhoods and areas in which people live. The researchers using the resource will work to answer the priority questions set by government and from the NHS. The findings will be fed back to help guide policy and health care decisions.

3. Project administrative details

3.1 Study Title:

- The Longitudinal Linkage Collaboration: a research platform for longitudinal COVID-19 research.

This is a major infrastructural component of:

- The Longitudinal Health & Wellbeing National Core Study

In turn, this is one of six research themes which, together with a 'Data & Connectivity' theme, comprise:

- The National Core Studies for COVID-19 research

3.2 Start Date: 01/10/2020

3.3 Duration: 3 years (with initial funding provided until 31.03.2021).

3.4 Institutions: The Longitudinal Linkage Collaboration is being established and is owned by the University of Bristol in collaboration with the University of Edinburgh. Swansea University and the University of Leicester are additional collaborating partners. A wider group of institutions and UK government departments, including the NHS, will supply data into the platform.

The wider Longitudinal Health & Wellbeing National Core Study is being led by University College London.

3.5 Funding Details: Initially supported by HM Treasury (until 31.03.201). It is then anticipated to be supported by UK Research & Innovation (UKRI). UKRI is the accountable body for the Longitudinal health & Wellbeing National Core Study and hence this study. The Medical Research Council is administering this award on behalf of UKRI. The grant code for the Longitudinal Health & Wellbeing study is: MC_PC_20030.

3.6 Project Approvals:

Health Research Authority Research Ethics Committee:

- This study has been submitted to the Health Research Authority's Haydock Research Ethics Committee. This protocol (Version 1) was reviewed on 10/11/2020 and was approved by the committee.

  The HRA REC review code for this study is: 20/NW/0446

  The IRAS application number for this study is: 290946

4. Objectives for the UK Longitudinal Linkage Collaboration

The UK has a world-leading portfolio of longitudinal population studies (LPS) which have collected detailed phenotypic and biological information on an estimated >3m members of the UK public across over 50 separate studies[1i]. These studies are ideally suited to answering certain research questions relating to the Severe Acute Respiratory Syndrome Coronavirus 2 (C-19) pandemic, given their detailed life-course information on health, wellbeing and behaviours. These LPS typically have extensive biobanks, including large scale participant genotyping and a data collection infrastructure that has been rapidly deployed to collect C-19 data. While there are increasing examples of cross-cohort research, of phenotypic and 'omic consortia studies and also study collectives (e.g. CLOSER, Dementias Platform UK), the studies have typically operated independently with distinct operating traditions and data collection strategies. The need to effectively respond to the C-19 pandemic has changed aspects of this. In response to the pandemic, the Wellcome Trust has convened a longitudinal group – led by University of Bristol and University of Edinburgh - to develop a standardised data collection survey[2] capturing participant symptoms, C-19 outcomes, participant behaviours and circumstances and wider health and wellbeing during the C-19 behavioural restrictions. These surveys are now being deployed by LPS and are being issued across multiple waves of collection with the survey instruments adjusting to the evolving pandemic and associated behavioural restrictions. The data from these will provide insights not available in routinely generated records[ii,iii,iv,v].

In order to optimise the LPS community's ability to contribute C-19 research insights the following requirements will need to be met:

(1) Urgent research questions related to the risk of C-19 and the consequences of the pandemic **requires linked and regularly refreshed electronic healthcare data alongside the existing longitudinal data** in order to conduct case ascertainment, to assess differential health outcomes and help-seeking behaviours;

(2) There will be many long-term consequences of the C-19 pandemic on multimorbidity and individual health conditions and the understanding of these **requires a long-term strategy for appropriate use of linked healthcare records in LPS for public benefit which is not encumbered by narrow definitions of C-19 research**. LPS research will likely have both health and socio-economic aspects and will consider the intersections between these (e.g. mental health outcomes resulting from financial anxieties;

(3) LPS are long-term studies operating over decades and built on a trust relationship with participants. Any expediated mechanism to enable C-19 research using participant health records will need to take a long view which **requires the safeguarding of participant rights, transparency in its activities and the accommodation of existing study-participant commitments**. The infrastructure should be co-designed with participant input. This will

[1] Medical Research Council. Strategic Review of the Largest UK Population Cohort Studies. Medical Research Council; London: 2014. www.mrc.ac.uk/populationcohortreview.
[2] See: https://www.bristol.ac.uk/alspac/researchers/wellcome-covid-19/ for further details.

help ensure the use of routine records retains 'social licence' [vi]and is sustainable going forward.

Access to C-19 relevant data should be delivered quickly for LPS to add value to C-19 understanding and inform emerging policy development. It should be implemented in a co-ordinated way to minimise burden on the NHS, linking standardised and existing datasets. Coordination across the C-19 research programme will bring efficiencies and best practice in common requirements (e.g. record integration processes, documentation, access and accreditation mechanisms).

This protocol describes the longitudinal linkage platform methodology for developing and implementing a data and governance pipeline to C-19 specific electronic health and administrative records and subsequently, for making the linked data available for C-19 research to legitimate users. This platform methodology has been co-developed with contributing studies and led by University of Bristol and the University of Edinburgh. The protocol is designed to be applicable to a wide range of LPS, but to be based around a central coordinating institution (a hub and spoke model). It builds upon the infrastructure for population data science being developed by the NHS, Health Data Research UK (HDR UK), Administrative Data Research UK (ADR UK) and other LPS infrastructure investments. The protocol will be iteratively developed with key stakeholders and subject to review following the C-19 emergency in line with other activity relating to the C-19 dataset. It will be shaped by participant and public views and representatives of the NHS's National Data Guardian. The data flows will be designed with consideration for the administrative, legal and jurisdictional boundaries of the devolved NHS systems; while considering options to leverage UK-wide insights through proportionate and acceptable ways of working.

Our Objective for the Longitudinal Linkage Collaboration's research platform:
Our primary objective is to co-locate and integrate a set of inter disciplinary LPS collected data within a secure centralised analysis platform and to then link these data with study participants' health and administrative records in a manner that is compatible with ethical and legal requirements, is publicly acceptable and is efficient to all parties.

The resulting dataset will enable C-19 research investigations across multiple LPS datasets. The following data will be used:

From contributing LPS

- Novel LPS C-19 questionnaire data, fieldworker assessed data or information assayed from biological samples collected during the pandemic (e.g. C-19 status from serology testing);
- Broad ranging 'baseline' LPS data on participant demographic, socio-economic and health status collected directly from participants or via linkage to routine records or novel data sources (such as social media posts, personal sensors) prior to the pandemic;
- Existing assayed biological information (e.g. Blood group type, specific DNA information, and potentially whole genome sequence data and other 'omic datasets);
- Participant consent/opt-out status;

From the NHS (separately in England, Scotland, Wales, NI):

- The centralised C-19 minimised datasets being developed by the devolved NHS authorities: to include primary care, secondary care, C-19 test data, the Shielded Patient List of clinically extremely vulnerable patients. This will include records on socially sensitive topics such as mental health and addiction;

From the other research studies:

- Symptom and outcome data from the C-19 Symptom Tracker app developed by Zoe Ltd and Twins UK (Kings College London);
- UK Biobank data: The LLC will submit an application to UK Biobank to access their C-19 datasets for the purpose of the work of the LH&W NCS.

Newly commissioned data:

- Geo-spatial resources (e.g. mapping tools and geographies, national geocoded resources and derived socio-economic and neighbourhood/land use indicators) and exposure estimates (e.g. air pollution exposure estimates);
- Contextual meta-data defining the C-19 context in which participants were living and reporting outcomes and behaviours (specific across time and location).

The integrated data will then need to be made available (onwardly shared) to legitimate research users: which will require the community to resolve pre-C-19 data governance challenges[vii]. A mechanism will be needed to promptly assess the legitimacy of the users and their research questions, this will need to meet study-specific obligations yet use national best practice frameworks (such as the Five Safe's approach) and the tools being developed for data discovery and access (i.e. HDR UK Gateway). All use will need to be transparent to the study participants, the wider public, and to contributing data owners. The research data used in investigations will need to be de-identified to protect participant confidentiality, and all outputs will need to be population-level research outputs that are checked for disclosure risk.

5. Protocol

This protocol will initially describe the methodology for the LLC and also the research programme for C-19 that will be conducted by the LH&W NCS.

In the first half of the protocol we will describe the contributing studies; the roles and responsibilities of study staff, LLC staff and then research users; the LLC infrastructure; the data flows coming into and out of the LLC and the methods by which these are regulated. We will then describe the process by which users and proposals are assessed and regulated through the research lifecycle.

In the second half of the protocol we will describe the data management and statistical methodological processing of the data and the proposed C-19 research programme.

5.1 Contributing Longitudinal Population Studies

The LLC is formed as a collaboration, with a central team based at the University of Bristol (home to the Avon Longitudinal Study of Parents and Children [ALSPAC] cohort) and University of Edinburgh

(home to the Generation Scotland cohort). The LLC is conceptually designed to be able to accommodate as broad and diverse a range of studies as is possible. In the first instance however, a 'vanguard' of 15 contributing studies has been identified (See Appendix 1). The scope of this protocol is currently limited to including these vanguard studies only.

The vanguard is interdisciplinary, has a mix of UK wide or regional/local catchments and has sampled individuals from across the life-course (currently aged from in utero to 100+). Over 1m participants are included across the sum total of the vanguard studies. Vanguard studies operate a cohort design except Understanding Society which is a longitudinal household panel study. The vanguard includes an existing cohort consortium - the HDR UK Multi-omics consortium – which will integrate and harmonise LPS collected data with assayed biological sample data in order to understand how variation in DNA impacts on health outcomes (including C-19 susceptibility and outcomes). While the LH&W NCS has an ambition to include the consortium and make use of its harmonised data set for C-19 research, primary contractual and governance mechanisms will be established with each of the contributing studies and their institutions (as Data Controllers) rather than with the consortium itself (which will facilitate)

These studies share three critical features: 1) they have collected C-19 data from their participants at least once during the pandemic; 2) they hold baseline data relevant to C-19 policy questions which was collected prior to the pandemic; 3) they have, or are able to, established the requisite permissions to flow their data into the LLC platform and enable it to be linked to at least some classes of routine health and social records.

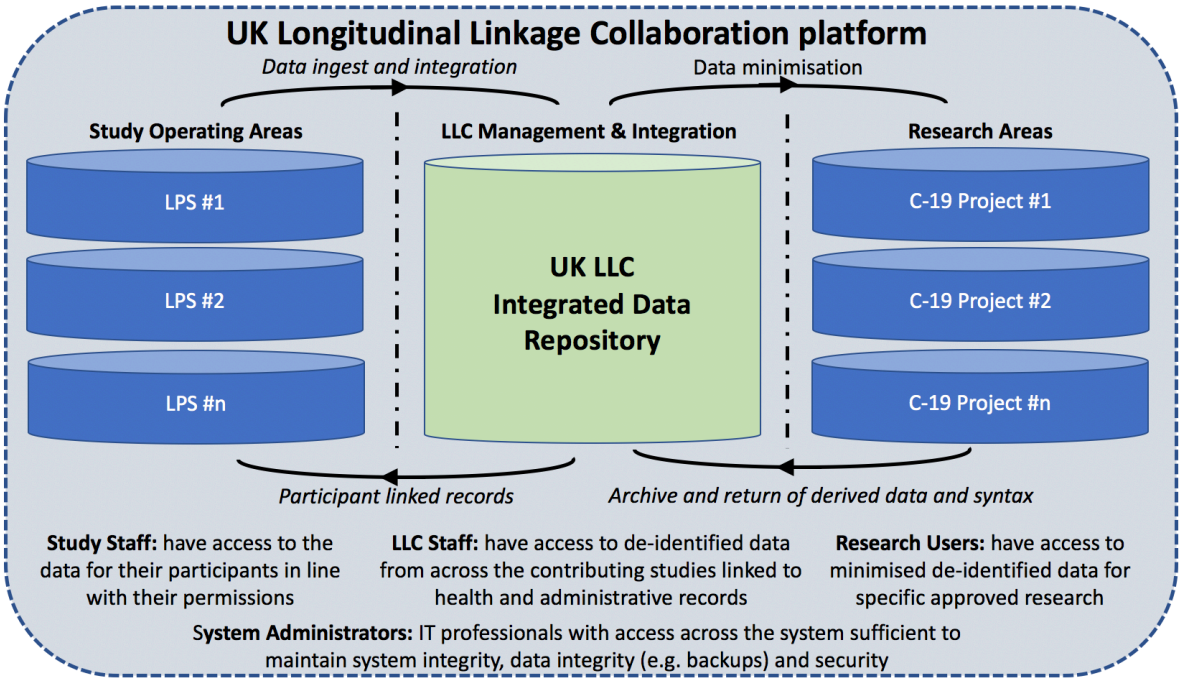
All studies will be expected to contribute C-19 relevant data (defined in 'LPS data flows' below) which comprise C-19 assessments, socio-economic and demographic data, pre-pandemic baseline assessments and assayed biological samples including genetic and other 'omic data. Data managers at contributing studies will also flow consent status on a quarterly basis into the LLC so that changes of consent can be implemented to stop linkage flows or research use of individuals' data.

5.2 Roles and Responsibilities for the different stakeholders within the LLC

There will be four distinct groups of users who have access rights to the LLC platform. These access rights will be based on distinct roles and data access will be minimised to each group (Figure 2):

1) *Study staff* from contributing studies who are responsible for and able to use data on their participants;
2) *LLC staff* who are responsible for managing the resource, integrating the data from studies, conducting quality assessments, deriving data products, provisioning data to research users and for conducting research;
3) *Research users* who are provided with minimised extracts of data sufficient for specific research projects;
4) *System administrators* who are responsible for maintaining system and data integrity.

Figure 2: Role based access to data within the LLC



Study staff (as appropriate within study governance structures) from each contributing study will have an operating area within the LLC where they can load and access their study data along with linked records for their study participants. Study staff will use this area for data management and processing and study specific research analysis (with prior LLC application and registration in the LLC governance systems). Study staff will not be able to access or alter the central LLC data resource.

The LLC staff will have access across the repository. This is necessary in order to be able to extract, transform and load (ETL) data from the LPS study areas into the repository and to feed study specific views of linked records back into the LPS study areas. The LLC staff also have the responsibility for providing minimised views of LPS data and linked records into the research project areas, for auditing the research process and for the archiving and return of derived variables, syntax and documentation created during the research process back into the primary LLC resource.

Research users will only be able to access a view of LLC data minimised to the needs of their specific project and where studies who object to the use of their data for a project are filtered out and where the rights of objecting participants are also respected. Research users will not be able to access or alter the central LLC data resource.

System administrators will be operating as Data Processors under contract to the UK LCC (University of Bristol).

5.3 Trusted Research Environments in UK health data research

A network of national Trusted Research Environments (TREs, also known as Data Safe Havens) are being established in the UK[3]. This is in response to drives for greater use of health care records in research and population health management coupled with extensive public consultation which has identified that the public are typically comfortable with the use of anonymised medical record data being used in research where this brings public benefits [viii]. This acceptance is conditional on sufficient safeguards being in place to protect individual's data during the research process: TREs are designed to meet, or facilitate, many of these safeguards. The UK Medical Research Council's Health Data Research UK has been leading the move to this new way of working through their UK Health Data Research Alliance. The C-19 pandemic has accelerated work in this area and the development of TREs and new data flows within TREs[ix]. The UK's Chief Scientific Advisor) recently made a clear case for the need for new data flows, with appropriate governance, to support the C-19 response (and for emergency situations more generally).[x]

TREs are complex constructs: comprising sufficient technical resources to maintain information security and provide efficient computing functionality, and also sufficient 'social' aspects – policies, procedures and ways-of-working – to ensure that they are ethically and legally compliant, transparent in their operations and aware of and responsive to changes in context.[xi] A TRE has three distinct elements: i) a *secure research infrastructure* – i.e. an IT system for secure analytics and which is managed to help ensure data integrity (e.g. the platform will have built in 'redundancy' which enables faults to be managed without disruption to the system, it will ensure backup management) and data security (e.g. firewall settings, data encryption, account management); ii) a *data management framework* which incorporates systems to host and organise data (e.g. a database 'warehouse', data models for organising and presenting data) and meta data (e.g. data discovery catalogues, application systems) and contextual data (e.g. health code look up libraries); iii) a *governance framework* of policies and procedures necessary to meet ethico-legal requirements and the requirements of key stakeholders (e.g. data owners, participants). Importantly, the *secure research infrastructure* layer and *governance framework* layer can be accredited to existing best practice standards (e.g. the ISO27001 Information Security standard, the NHS Data Security Protection Toolkit, The UK Statistics Authority's 'Accredited Research Environment'[4]).

The TRE approach is seen as offering the following benefits:

1) Rather than providing extracts of data to researchers (a 'lending library' approach) the TRE provides a secure analytics environment in which research can take place (a 'reading library' approach);

[3] With NHS Digital establishing a TRE for C-19 research in England (https://digital.nhs.uk/coronavirus/coronavirus-data-services-updates/trusted-research-environment-service-for-england); A long-standing collaboration between Public Health Scotland, National Records for Scotland and Edinburgh University providing a TRE for C-19 research in Scotland (https://www.epcc.ed.ac.uk/blog/2020/06/our-response-covid-19); and the SAIL Databank with NHS Wales and the NHS Wales Informatics Service providing a TRE for C-19 research in Wales (https://saildatabank.com/).
[4] https://uksa.statisticsauthority.gov.uk/digitaleconomyact-research-statistics/better-access-to-data-for-research-information-for-processors/

2) TREs are designed to fulfil the requirements of the Information Commissioner's Anonymisation code of practice[xii] which provides a template for applying sufficient controls on the data that the risk of disclosure is no longer reasonably likely and can be considered as being effectively anonymous whilst the controls are in place (and therefore that the data are no longer Personal Data);
3) A TRE applies safeguard controls to both the data (e.g. removing identifiers, encrypting research relevant information – such as postcode – so it is no longer identifiable, output disclosure risk checks) and the context in which the data are processed (e.g. user contracts, training, independent audits, infrastructure IT settings). This means that a balance can be struck so as to limit the degradation of data through anonymisation processes and the impact this can have on research findings;
4) TREs are also compatible with the principles set out in the Digital Economy Act (2017) which provides a means of flowing and using administrative (non-health) data for public benefit research;
5) TREs provide a mechanism for enabling 'five safes' research: where the research user, project, setting, data and output are assessed as being demonstrably 'safe' against a set of consistent criteria (as set out by the UK Statistics Authority);
6) Through designing and auditing TREs to a set rigorous standard, this can generate public trust in health data research through applying meaningful safeguards and clear and consistent messaging;
7) Being able to meet public expectations for the safe use of data.

There are precedents for the use of TREs in longitudinal research. Some biomedical LPS have established their own independent TRE[5], whilst others have sought to do this via research consortium projects[6]. In contrast to this, the Economic and Social Research Council support a central infrastructure called the UK Data Service - which includes TRE functionality - to hold, process and share the data from the LPS they fund[7].

The primary driver for LPS to establish these TREs has been the need to meet and demonstrate higher security and governance requirements in order to support linkage to NHS and administrative routine records. However, most LPS do not have TREs in place, and many smaller LPS do not have the resources to develop and maintain a bespoke TRE; which introduces barriers to linkage programmes within the LPS community. Not all existing TREs have the functionality to meet the needs of all LPS (for example, the ethico-legal requirement for data minimisation which is seen as prerequisite for sharing biomedical data is currently not fully supported within the UK Data Service). Diverse governance needs based on study specific ways of working and past assurances to

[5] For example, the ALSPAC birth cohort study operate an ISO27001 and NHS Data Security Protection Toolkit certified TRE based on UK Secure eResearch Platform infrastructure and a bespoke governance layer developed by the study. This approach incorporates participant requirements and also meets data owner expectations and those from ethical review panels.
[6] For example, the Dementia Platform UK and the HDRUK BREATHE HUB LPS consortium studies are both based on UK Secure eResearch Platform where they host data sourced from across multiple LPS.
[7] UK Data Service at the University of Essex provides a secure TRE certified to ISO27001 and UK Statistics Authority standards (https://www.ukdataservice.ac.uk).

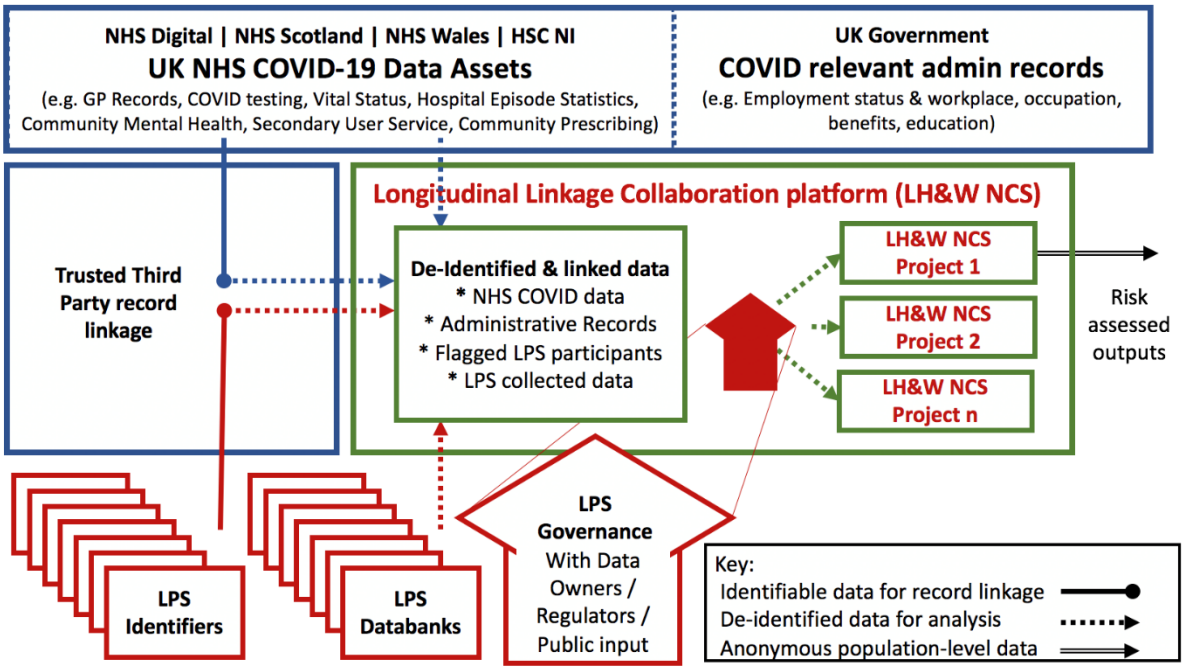
participants describing the ways in which data sharing would be conducted has also been a barrier to establishing TREs.

The combination of these factors makes an extremely compelling case to coordinate linkage by the UK LLC within a centralised TRE. However, managing diverse governance and ethico-legal basis of the different LPS into one single governance framework would be complex. The most feasible option is to use an existing and accredited infrastructure where the maintenance and evolution of the underlying *secure research infrastructure* is managed by specialist staff and funded centrally by UK research funders (although supported by the UK LLC through cost recovery mechanisms) and then for the LLC to co-develop with contributing studies a new *data management process* and a *governance framework* which meets the collective needs of all contributing studies.

5.4 The Longitudinal Linkage Collaboration TRE

The LLC TRE (Figure 2) will provide the platform for studies to upload data, to establish record linkages and to extract the routine records of participants, for LLC staff to conduct data management and integration (across studies and linked NHS and other records) and then to provide research users with separately partitioned areas for research analysis. All analysis of LLC data will take place within the TRE and only aggregate outputs checked for disclosure risk will be allowed to exit the system.

Figure 2: COVID-19 dataflows into and out of an LPS enriched TRE



5.5 The use of the UK Secure eResearch Platform to support the LLC TRE

The UK LLC will contract University of Swansea to provide a UK Secure eResearch Platform (UKSeRP) environment as the basis on which to build the LLC TRE. This contract will form a component of an academic collaboration which will also involve sharing the expertise gained from developing and

operating the Secure Anonymised Information Linkage (SAIL) Databank (itself based on a UKSeRP) to aid the development of the LLC TRE.

UKSeRP has been designed to provide secure research computing facilities for data science[8], including LPS[9]. The UKSeRP provides the secure infrastructure for the SAIL databank in Wales and hosts individual LPS such as ALSPAC[10], Generation Scotland and the Millennium Cohort Study[11], and LPS consortia initiatives such as Dementias Platform UK[12]. UKSeRP is seen as the ideal TRE infrastructure to support the LPS COVID-19 data resource given that it:

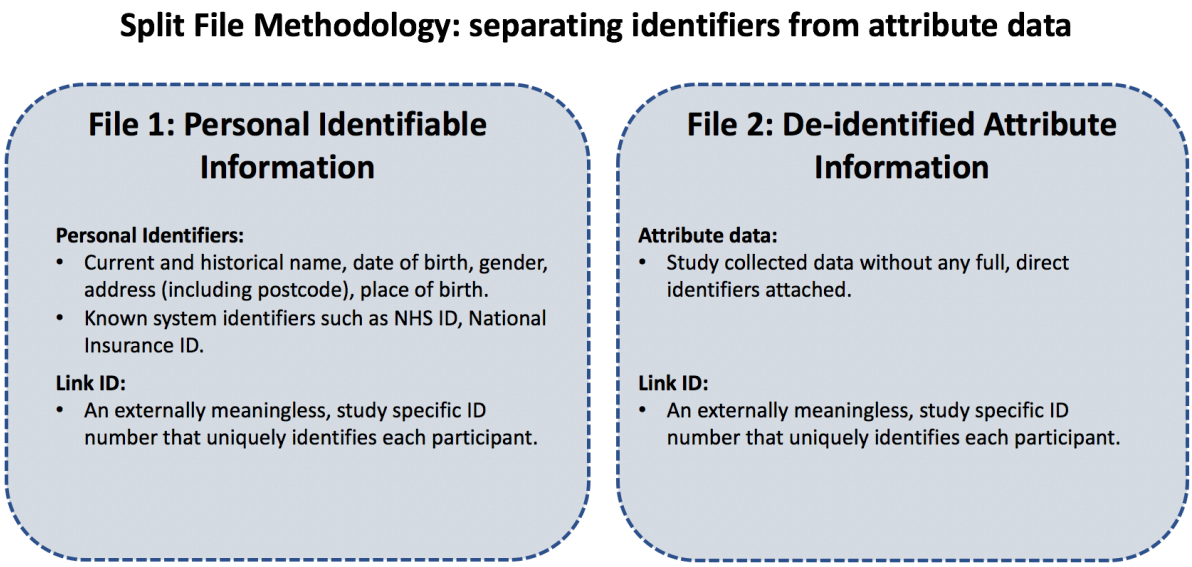
- It is recognised by key stakeholders as having a robust and independently accredited Information Security systems (UKSeRP is ISO27001 certified and has NHS DSPT certification) and is well regarded in its ability to host LPS and NHS data[13];
- Its linkage capabilities enable the linkage of NHS records and other sources with diverse identifiers in a privacy-preserving manner given they are conducted by a Trusted Third Party who do not have access to any individual attribute data (beyond individual identifiers);
- It is a fully de-identified environment and hosted data is effectively anonymous to all research analysts and the system administrators. This has security and governance advantages (to meet our DPA 2018 requirements) and is also consistent with study reassurances to participants where some studies have made assurances that record linkage is not a means by which identifiable participant data is transferred to the state;
- UKSeRP is capable of federated access with the Scottish TRE and is also providing infrastructure to support data science within Northern Ireland.

5.6 Methodology for the flow of data within the LLC TRE

The LLC TRE will be operated and managed by the LLC team (based at University of Bristol and University of Edinburgh). Contributing LPS will provide data to the LLC TRE using a 'split file' approach which is used across UKSeRP (Figure 3). In this methodology, the LPS data will be split by the LPS data managers into a file of personal identifiers (File 1) and an externally meaningless 'Link ID'. Separately, the attribute data (File 2) will be de-identified (direct and pseudo-identifiers will either be dropped or transformed into less identifiable research variables) and indexed using the same 'Link ID' as File 1.

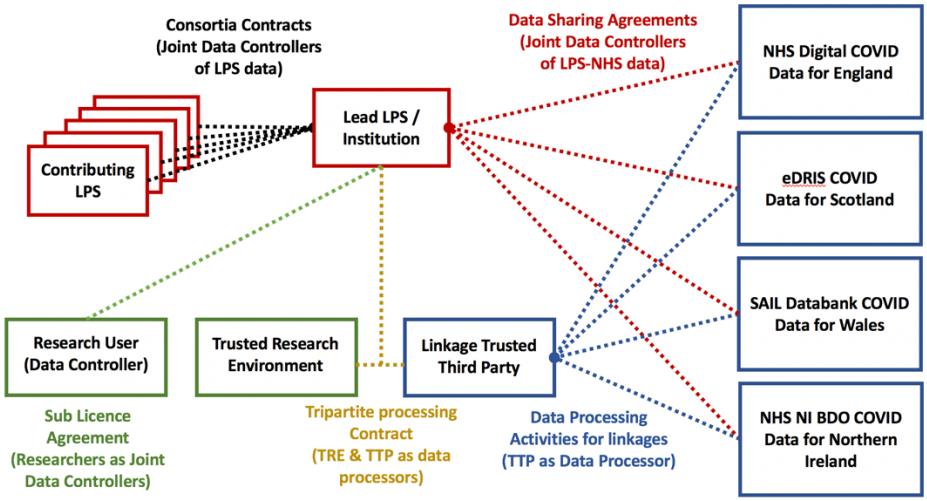
[8] Jones KH, Ford DV, Ellwood-Thompson S, Lyons RA. The UK Secure eResearch Platform for public health research: a case study. The Lancet. 2016 Nov 1;388:S62.
[9] Jones KH, Ford DV, Ellwood-Thompson S, Lyons RA. The UK Secure eResearch Platform for public health research: a case study. The Lancet. 2016 Nov 1;388:S62.
[10] Cornish RP, John A, Boyd A, Tilling K, Macleod J. Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R. BMJ open. 2016 Dec 1;6(12):e013167.
[11] Millennium cohort study data for Welsh residents has been deposited in the SAIL databank and linked to Welsh resident records (https://data.ukserp.ac.uk/Asset/View/52).
[12] Bauermeister S, Orton C, Thompson S, Barker R, Bauermeister J, Ben-Shlomo Y, Brayne C, Burn D, Campbell A, Calvin C, Chandran S. Data Resource Profile: The Dementias Platform UK (DPUK) Data Portal. BioRxiv. 2019 Jan 1:582155.
[13] NHS Digital Audit of Data Sharing Activities: University of Bristol, ALSPAC. (2019) NHS Digital. Leeds, UK. Available from: https://digital.nhs.uk/binaries/content/assets/website-assets/services/dars/data-sharing-agreement-audit---university-of-bristol.pdf

Figure 3: 'Split file' approach to separating identifiers from attribute data



The LPS will encrypt[14] and send File 1 to the LLC Trusted Third Party (TTP) for linkage. This linkage service will be conducted by the NHS Wales Informatics Service who will act as a linkage 'broker' and facilitate linkages across the four NHS authorities. The NWIS TTP will then forward this File 1 to the relevant NHS organisations for linkage across the UK Nations (NHS Digital in England, Public Health Scotland in Scotland, NWIS in Wales, NI BDO in Northern Ireland). The LLC will facilitate the linkage through establishing the contracts and data sharing agreements needed to permit this (figure 4) – but will not handle the study participant identifiers (File 1s) at any stage.

Figure 4. Contractual arrangements supporting the LLC.

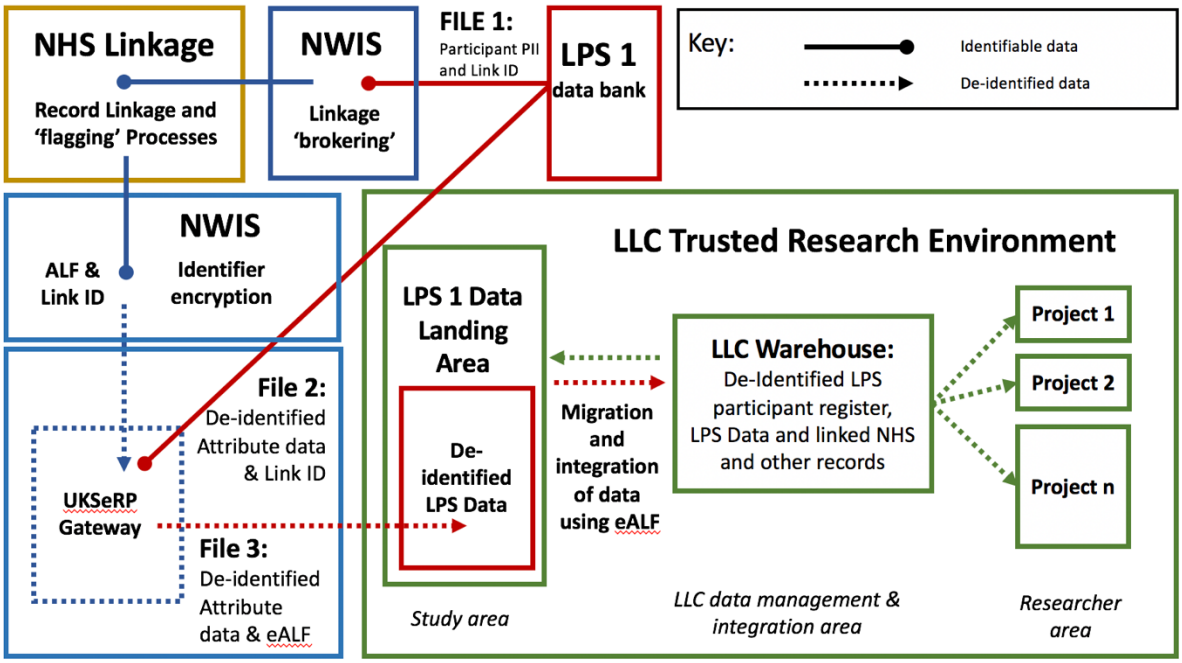


Many UK LPS will have established linkages between their administrative study databases and some of the UK NHS patient registers, resulting in study membership being 'flagged' on the patient

[14] sending the decryption password by a different mechanism

demographic record. Where this flagging exists, it will be used to extract and provide records into the LLC via the TTP. Where study flagging does not exist, the relevant NHS organisation will use the File 1 identifiers to link the LPS participants to their patient register[15]. This flagging will then be a persistent asset which can be reused to enable data flows over time. Maintenance of the flagging will be subject to contributing LPS flowing dissent status (via a File 1 and the TPP) to remove flags for participants withdrawing from the study or dissenting from the use of NHS data in their research. NWIS as TPP will then arrange for the NHS organisation to remove the identifiers from File 1 and replace these with a consistent 'Anonymous Linkage Field' ID. The NHS organisation will then flow this into the UKSeRP (University of Swansea) 'gateway' with Link ID (Figure 5). Separately, the LPS will send their File 2 (de-identified attribute data with Link ID) to the gateway. An automated process will then join File 1 with File 2 and remove Link ID: this process produces File 3 which contains ALF and the attribute data. ALF in File 3 will then be encrypted a second time into eALF which is unique to the LLC.

Figure 5: detailed data flows illustrating linkage between LPS and the UK LLC



The same mechanism will be used to integrate other data sources: for example, NHS Digital could send a File 1 to the gateway (via the TTP) and a separate File 2 with attribute data. The ALF encryption and re-encryption process is consistent across different data sources, nations and time. Meaning that the eALF identifying data entering the LLC TRE can be used for internal linkages (e.g. joining LPS collected data with the participants NHS records) and also for compiling an LLC de-identified participant register as eALF consistently identifies a unique individual.

[15] In England, NHS Digital will flag participants onto the Patient Demographic Service (PDS) patient register; in Scotland PHS will flag participants onto the Scottish NHS register; in Wales NWIS will flag participants onto the Welsh Demographics System (WDS). The mechanism for linkage in NI is yet to be clarified and is outside the scope of this protocol.

Geospatial linkages

This mechanism also works for geospatial data. In this instance 'File 1' will refer to a location (e.g. a property) which is also assigned a Link ID. This is linked to a register of all locations (in this case properties) and the location value (property ID) is encrypted into a RALF (Residential Anonymous Linkage Field) which will then have secondary encryption eRALF before entering the LLC. File 2 will contain environmental indicators and also Link ID. The geo-spatial coordination (e.g. measuring the distance between home and the nearest GP surgery) will be done outside of the LLC and only the derived variable (checked for disclosure risk) is allowed into the LLC identified and linked to individuals via RALF. This process enables environmental exposures and characteristics to be quantified precisely, and linked to participants, without directly linking potentially disclosive location information with LPS/health data and therefore maintaining confidentiality[xiii].

The data flowing into the LLC will be processed, integrated and managed by LLC staff. The LLC staff will produce minimised sub-sets of data for each approved research project and make these available for research (within demarcated and access-controlled working areas). Each project will have its own re-encrypted eALF which is specific to that project.

Separately, LPS staff will have access – within their own study specific demarcated area – to the data (both LPS collected data and linked NHS records) for their participants. This will be indexed using a re-encrypted eALF which is specific to that study. The study staff will not have access to the key to reverse eALF back to their own participant identifiers. Study staff will not have access to the data for participants in other LPS or collected by other LPS.

The outputs from the LLC TRE project areas will be fully anonymous population-level research outputs which are checked by for disclosure risk.

5.6 A de-identified participant register and interactions across study samples

We will need to address some new data challenges relating to the pooling of multiple large LPS into a single environment. It will be necessary, using encrypted de-identified information from the linkage TTP, to define individual people and properties across the sum total of contributing studies in order to create a de-identified participant register. This is crucial as the joint analysis of data across multiple studies will be conducted on a statistical assumption that the samples are independent: yet it is known that this is not the case. For example, a linkage exercise identified that almost 1,000 of the original ALSPAC mothers are also participants of UK Biobank (approximately 6% of the enrolled sample). It is also important in the context of COVID-19 research to define household occupants (for example, which individuals across multiple generations are living in the same property).

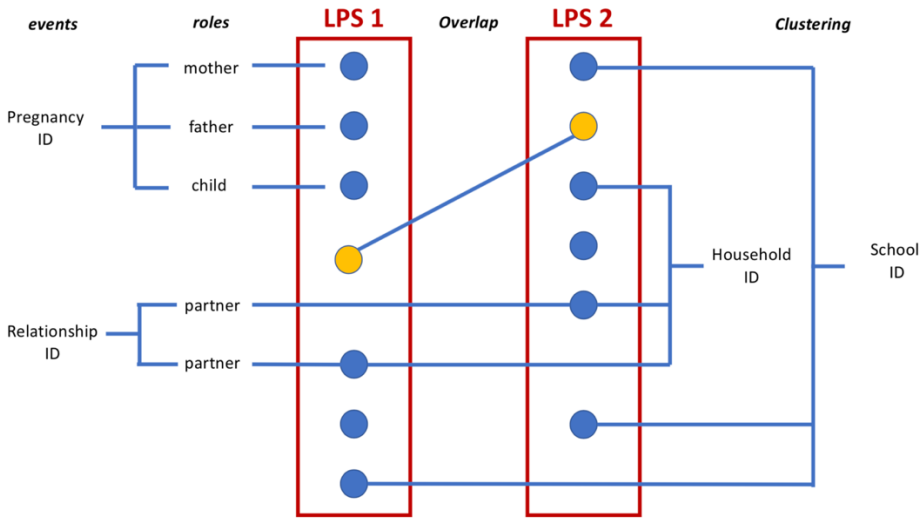
The register will need to capture information across multiple dimensions (see Figure 6):

- **Identifying unique individuals**: including the *overlap* of individuals across multiple contributing studies and where individual identity can be somewhat challenging to

consistently determine (e.g. birth order amongst multiple births, where individuals have changed identities);

- **Annotating the roles/relationships between individuals**: within studies it is common to track and annotate the roles and relationships between individuals, for example a mother-child pair, or that individuals are partners;
- **Identifying clustering of individuals at household and other levels:** it will be important for some COVID-19 research to identify household composition and the temporal variation in this (for example, students or renters returning to the family home during lockdown and a period of financial stress). Household composition will be determined through using encrypted de-identified property reference data (unique property ID) provided by the linkage TPP. This principle could be extended to other levels, for example neighbourhoods, schools, care homes or workplaces through de-identified information linked through routine records.

Figure 6: Illustrative examples of potential complexities of events, roles, overlaps and clusters within a matrix of LPS samples.



The overlap of participant samples may introduce consent/permission ambiguities where permissions to link to and use routine records across different studies are set in contrasting ways. For example, an individual who is a participant of both LPS 1 and LPS 2 may have been informed by both studies about the studies use of linked health records in research and may have consented to LPS 1 and dissented to LPS 2. Where contradictions such as these exist, we will filter data extracts, flows and use according to a rule set (Table 1) designed to minimise the risk of participant harms. These filters will only apply when data are used across multiple studies and where inconsistencies exist. Where only one study is used or multiple studies with no contradictory permissions are used then no filtering will take place. I.e. in the example above, routine health records of the individual would be extracted and used for research based on LPS 1 but not for research including both LPS 1 and LPS 2. These principles will be scaled to apply to any situations where participants are members of three or more studies (it is not known whether any participants will fit this scenario).

Table 1: Implementing participant level permissions for the use of routine records across multiple studies: a permissions matrix illustrating which participants are to be included in data use and which should be excluded.

			LPS 1 *(Study linkage permissions methodology)* Participant permissions				
			(Opt-in Approach)		*(Opt-out Approach)*		*(No Approach)*
			Consent	*Objection*	*No Response*	*Objection*	*N/A*
LPS 2	*(Opt-In Approach)*	*Consent*	*Included*	*Excluded*	*Included*	*Excluded*	*Excluded*
		Objection	*Excluded*	*Excluded*	*Excluded*	*Excluded*	*Excluded*
	(Opt-Out Approach)	*No response*	*Included*	*Excluded*	*Included*	*Excluded*	*Excluded*
		Objection	*Excluded*	*Excluded*	*Excluded*	*Excluded*	*Excluded*
	(No approach)	*N/A*	*Excluded*	*Excluded*	*Excluded*	*Excluded*	*Excluded*

6. Legal Basis

6.1 Legal basis for the LLC

The LLC is owned by the University of Bristol. The University of Bristol enacting legislation provides a lawful basis to conduct research. LLC will not process Personal Data so is not required to have a legal basis under EU General Data Protection Regulations and Data Protection Act (DPA) 2018 and its activities will not breach Common Law Duty of Confidentiality.

6.2 Legal basis for flowing LPS data

All contributing LPS are owned by Universities except for Born In Bradford which is owned by the Bradford Teaching Hospitals NHS Foundation Trust (BTHFT). The Universities owning the LPS and BTHFT have enabling legislation which includes a remit and lawful basis to conduct research. The contributing studies legal basis under GDPR and the Data Protection Act 2018 will be 1) performance of a task carried out in the public interest (Article 6(1)(e) in the GDPR); and, for the use of sensitive personal information, 2) scientific or historical research purposes or statistical purposes (Article 9(2)(j) in accordance with Article 89(1)). The studies will meet Common Law Duty of Confidentiality either through explicit consent, consent exemptions under the Health Service (Control of Patient Information) Regulations 2002 within England and Wales (Regulation 5 with support from the Health Research Authorities Confidentiality Advisory Group) or through undergoing public interest test assessments by the relevant UK devolved authorities (through the Public Benefit and Privacy Panel in Scotland).

6.3 Legal basis for the flowing of NHS data from England into the LLC

The Health and Social Care Act (2012) provides a statutory basis for the sharing of health data in England. We will consult with NHS Digital regarding flowing English residents NHS records into the LLC. NHS Digital's basis for flowing data under the DPA will be 1) performance of a task carried out in the public interest (Article 6(1)(e) in the GDPR); and, for the use of sensitive personal information, 2) scientific or historical research purposes or statistical purposes (Article 9(2)(j) in accordance with Article 89(1).

6.4 Legal basis for the flowing of NHS data from Scotland into the LLC

The LLC will submit an application to the Scottish Public Benefit and Privacy Panel (PBPP) to request the flow of data to the LLC. For those LPS who use consent as the basis of the data flow there are already precedents in place for this approval (e.g. UK Biobank). Those LPS who rely solely on Section 251 approval in England may not be able to be included in this data flow. Public Health Scotland's basis for flowing data under the DPA will be 1) performance of a task carried out in the public interest (Article 6(1)(e) in the GDPR); and, for the use of sensitive personal information, 2) scientific or historical research purposes or statistical purposes (Article 9(2)(j) in accordance with Article 89(1).

6.5 Legal basis for the flowing of NHS data from Scotland into the LLC

NHS Wales has provided de-identified data into the SAIL Databank in such a way that it is not Personal Data whilst in the protective controls of the TRE. The SAIL databank is permitted to flow these de-identified data across UKSeRP platforms and therefore into the LLC UKSeRP.

6.6 Legal basis for the flowing of NHS data from Northern Ireland into the LLC

It is our intention to develop data flows for NHS NI records into the UK LLC. This work remains under discussion and is not included within this version of our protocol.

7. Data Flows

7.1 LPS Data Flows

COVID-19 relevant data (table 2) will be deposited into the LLC platform by data managers from participating studies using the 'split file' protocol. The data will be pre-processed so that it is cleaned to study standards and direct identifiers removed. The study data will be indexed using a pseudonymised ID number which can inform the TTP linkage and ID encryption process.

Due to the quick timeframes required by funding to get the LLC up and running, Generation Scotland (GS) will provide one data flow of their linked health care data that is C-19 relevant to the LLC. We are prioritising linkage to NHS England data in the first instance as all the LPS with the exception of GS are primarily English. The application for data from Public Health Scotland will be submitted by

the end of the initial funding period (31.03.2021), but we do not expect data to flow from Scotland into the LLC until the second quarter of 2021 at the earliest.

Table 2: Study COVID-19 relevant datasets

Category	Collection Methodology	Example Contents
COVID-19 Assessments	Questionnaires	COVID-19 symptom, diagnosis and outcome information; information about the home and other occupants; mental health status, behaviours (including exercise; diet; alcohol, smoking and drug use; gambling); social distancing behaviours; home schooling patterns and outcomes; use of services and neighbourhood/countryside (green and blue space); neighbourhood conditions.
	Fieldworker assessments	Direct follow-up of participant status through face-to-face or remote assessments.
	Assayed samples	E.g. Immunology and serology test outcomes
Socio-economic and demographic indicators	Date of birth	Participant date of birth (which is needed at this precision for data management to enable the sequencing of information; but will be provided to researchers in age in years (or days/weeks/months for neonates to 1 year).
	Gender	Participant gender
	Ethnicity	Participant ethnicity
	Socio-economic indicators	E.g. Housing tenure; education levels (including maternal/paternal education); occupation (categorised); earnings (categorised); council tax band.
	Neighbourhood indicators	E.g. Indices of Multiple Deprivation (deciles); urban/rural status.
Baseline data	Questionnaires & Fieldworker assessments	Any pre-pandemic baseline data of direct relevance to approved COVID-19 investigations (to include socially sensitive subjects including mental and sexual health status, addictions, domestic abuse).
	Assayed samples	Genetic sequence information, Blood group, other 'omic information

7.2 NHS Data Flows

To be fit for LPS purpose, and to meet our objectives, the NHS data sets should have *maximum population coverage*. It is important that (wherever possible) filters are applied after extraction as the LPS will be in some cases be consent based (thus overriding National Opt Out). It will be important to extract individuals' *full life-course record* as this will enable both the accurate ascertainment of pre-existing disease status and how preclinical disease and their trajectories impact on COVID-19 risk. The extract will need to be refreshed on a timely basis to ensure an accurate assessment of disease (COVID) status and outcomes, and on information about broader health status and health service interactions during the pandemic. The data collection specification used across the UK nations should ideally be prospectively aligned so enable efficient data

integration. Finally, consideration should be made as to the likely duration of such a resource: with a presumption that it will be used for short-term insights to inform policy development and service provision, and a more sustained use to investigate long-term outcomes and policy response following social restrictions and the outbreak.

The intention is to extract as full coverage of our participants from the English population as possible. However, it is recognised that each contributing dataset will have its own inclusion/exclusion criteria, that some cases will be excluded from some data through setting patient objections (e.g. the primary care extract will respect Type 1 patient objections which block the sharing of records for purposes other than direct care), and that the NHS system has protective mechanisms to restrict the sharing of patient health records in certain sensitive situations. This may mean that some vulnerable and marginalised groups are systematically excluded[16].

NHS COVID-19 data flows in England

As described in the HDRUK paper, NHS Digital will coordinate the development of extraction specifications through Data Provision Notices (DPNs).[17] These are envisaged to contain detailed life-course health records (Table 3).

Table 3: NHS COVID-19 relevant Data Sources in England*

Owner	Data Set	Contents
Public Health England	SGSS	COVID-19 lab test records
	CHESS	COVID-19 Hospitalisation in England Surveillance System records
Serology data	Serology test outcomes	Outcomes from citizen serology testing
NHS Digital	PDS	Vital status, demographic and area data, National Opt Out status
	HES/SUS	Hospital Episode Statistics
	Death Registration	Digitised death certificate information
	Community Prescription	Prescription records
	Primary Care	COVID-19 minimised extract of English GP records (COVID status and outcomes, mental health status, help seeking information).
	Shielded patient list	
ICNARC	ICNARC	Intensive Care national audit records
Disease Specific Datasets	HQIP & Vascular Datasets	NICOR, SSNAP, Vascular Procedure Registers
	HQIP & Other disease datasets	Cancer Registration, diabetes, renal, COPD etc

[16] Boyd A, Thomas R, Macleod J. (2018). NHS Number and the systems used to manage them: an overview for research users. Bristol, UK: University of Bristol. Available from: https://www.closer.ac.uk/wp-content/uploads/CLOSER-NHS-ID-Resource-Report-Apr2018.pdf

[17] NHS Digital issue a Data Provision Notice when they use their statutory power under section 259(1) of the Health and Social Care Act 2012 (the Act). https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/data-provision-notices-dpns

* Adapted from "A national health data research capability to support COVID-19 research questions. 15 April 2020. HDRUK. London, UK." SGSS: Second Generation Surveillance System; CHESS: COVID-19 Hospitalisation in England Surveillance System (CHESS); HES: Hospital Episode Statistics; SUS: Secondary Uses Service; BSA: Business Services Authority; PDS: Personal Demographic Service; ICNARC: Intensive Care National Audit and Research Centre; HQIP: Health Quality Improvement Programme; NICOR: National Institute for Cardiac Outcomes Research (datasets include Myocardial Infarction National Audit Project; Percutaneous Coronary Intervention audit; Cardiac Surgery audit; Heart Failure audit; Cardiac Rhythm audit; Congenital Heart Disease audit; Left Atrial Appendage Occlusion audit; Percutaneous Mitral Valve Leaflet Repair audit; Transcutaneous Aortic Valve Implantation audit; Patent Foramen Ovale closure audit); SSNAP: Sentinel Stroke National Audit Programme); COPD: chronic obstructive pulmonary disease; ISARIC-CCP: International Severe Acute Respiratory and emerging Infection Consortium – Clinical Characterisation Protocol.

NHS COVID-19 data flows in Scotland

Public Health Scotland provide access to Scottish health records in conjunction with National Records Scotland for registry data and Albasoft Ltd for primary care records (Table 4).

Table 4: NHS COVID-19 relevant Data Sources in Scotland

Owner	Data Set	Contents
Public Health Scotland (PHS)	Scottish Morbidity Records	• Outpatient Appointments and Attendances (SMR00) • General Acute Inpatient and Day Case (SMR01) • Maternity Inpatient and Day Case (SMR02) • Mental Health Inpatient and Day Case (SMR04) • Scottish Cancer Registry (SMR06)
	Prescribing Information System (PIS)	Data on all medicines and their costs that are prescribed and dispensed in the community in Scotland.
	Unscheduled Care	Data from: • Accident and Emergency • GP Out of Hours • Scottish Ambulance Service • NHS 24
	ECOSS	Electronic Communication of Surveillance in Scotland (ECOSS): Holds all positive microbiology laboratory specimen results and a subset of antimicrobial susceptibility/resistance data in Scotland
	SICSAG	The Scottish Intensive Care Society Audit Group (SICSAG) a national database of patients admitted to adult general Intensive Care Units (ICU) in Scotland

	MIDAS	Dental data
	Disease specific datasets	SCI-Diabetes, Scottish Stroke Care Audit, Scottish Renal Registry etc
National Records Scotland (can apply for access via PHS process)	Deaths	All Registrations to the National Records of Scotland of deaths
	Births	All Registrations to the National Records of Scotland of live births
Serology data	Serology test outcomes	Outcomes from citizen serology testing
PHS/Albasoft	Primary Care	COVID-19 related research only

NHS COVID-19 data flows in Wales

Welsh NHS records for Welsh residents are centralised and stored within the Secure Anonymised Infrastructure for Linkage (SAIL) databank in partnership with NHS Wales Informatics Service (Table 5). The SAIL databank also hosts records for the Symptom Tracker App (https://covid.joinzoe.com/) which has participant reported data on C-19 symptoms and status from over 4m individuals. The app was co-developed by the LPS community (Twins UK, Kings College London) and a de-identified copy of the data are stored in SAIL for population health research into C-19.

Table 5: NHS COVID-19 relevant Data Sources in Wales

Owner	Data Set	Contents
SAIL Databank / NHS Wales Informatics Service	Welsh Secondary Care records	• Critical Care Dataset • Emergency Department Dataset • Maternity Inpatient and Day Case (SMR02) • Outpatient Datasets & Referrals • Inpatients (Patient Episode Dataset for Wales) • Postponed Admitted Procedures • Maternity indicators dataset • Welsh Cancer Intelligence & Surveillance Dataset • Referral to treatment times dataset
	COVID-19 Shielded People List	A list of people at “high risk” of C-19 complications.
	COVID-19 Pathology test results	Test results from C-19 PCR testing.
	Care Homes	Residential and geographical information about care homes
	Substance Misuse Dataset	Help seeking for substance misuse dataset
Serology data	Serology test outcomes	Outcomes from citizen serology testing

SAIL Databank/Zoe Ltd.	Symptom Tracker App	data derived from the participant information provided through the use of a nationally-used smartphone app for collecting COVID symptoms
Welsh GPs	Primary Care	Primary Care record

NHS COVID-19 data flows in NI

It will be our intention to develop governance approvals to allow linkage to NHS NI records and the flow of these into the UK LLC. This work remains under discussion and is not included within this version of our protocol.

7.3 Administrative record flows

There is an ambition to incorporate administrative records into the LLC given the importance of these on the research question (e.g. the role occupation as a risk factor for C-19 exposure, status and outcomes; employment status and benefit provision on anxiety and mental health; home schooling on educational attainment and wellbeing). The mechanism for linking administrative records into the LLC is not yet clear so these data flows are not included in this version of the protocol. We do note however that these data are likely to flow using the Digital Economy Act (2017) (DEA) as a legal basis: and that the DEA data flow mechanisms utilise a 'split file' approach so which could be accommodated into a UKSeRP based TRE without disruption to existing data flows. Further to this, UKSeRP is already certified by the ONS/UK Statistics Authority as being an accredited research analysis platform and that the SAIL databank governance framework provides a template for an accredited TRE capable of hosting data which has flowed using the DEA.

8. Information Governance

8.1 Social Licence and LPS oversight

The trust relationship between a study and its participants is an integral component of the study's 'social licence' for sustainable ongoing participation[18]. Participant feedback stresses the 'trust' in this relationship is formed between the participant and the core study staff who act as custodians of the donated data. LPS managers will need to demonstrate to participants how they are custodians within this new resource. The framing and 'rules' relating to this will be specific to any given LPS: as all have a different history, were established for different purposes, at different times, have different sample composition and have provided a range of different assurances to participants over time. Some of these assurances will sit in the 'fair processing' information provided over time, and some studies may have provided more formalised social contracts (e.g. for ALSPAC: http://www.bristol.ac.uk/alspac/participants/our-commitment-to-you/). Given this variation across LPS, it may be challenging or impossible to reconcile without resorting to a lowest common denominator which would jeopardise the potential of the resource to meet the objective: we may establish 'tiers' of involvement which individual LPS can select based on their specific requirements.

[18] Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care. data ran into trouble. Journal of medical ethics. 2015 May 1;41(5):404-9.

Across the community, at a minimum, for the most permissive tier of involvement, the following will need to apply:

- Participant consent/objections are implemented at a study level. It is likely that there will be overlap between study membership (i.e. individuals are members of multiple studies), in which case a rule matrix has been established to determine how differing status should be interpreted;
- That all use of the resource is known, documented and made transparent to the study managers and study participants and the wider public. That risks are assessed, documented and acted upon (i.e. proportionate mitigations are put in place);
- That the purpose of the data use must be for generating improvements to the public good and not be for profit.
- That the data in the resource will be de-identified and that the end users will be accessing them on an 'effectively anonymous' basis. Outputs from the TRE will be at an aggregated population level and will have been reviewed for disclosure risk before leaving the TRE.
- LPS retain oversight of data flows and account permissions in order to effectively audit the data they are responsible for.

To account for variation in LPS local arrangements, the following may need to apply:

- LPS retain oversight of each use of the resource and can implement a veto as to whether their data is used or not (based on their specific data reuse rules): however the default principle is for the resource to be open (in a managed way) to the legitimate research community.

8.2 Participant fair processing and rights

It is essential that the data flows and use described in this protocol fall within participants 'reasonable expectations' as to how their data are being used. To ensure this, 'fair processing' information will be provided to participants in order to ensure these data flows and research aims are understood. This will be accompanied by a clear route to object if this is not considered acceptable. The fair processing materials will be summarised at a consortium level, but finalised and implemented at a study level (who have the best understanding of their cohort and the existing information that has been provided). The materials will be guided by best practice (e.g. the insights generated by the Understanding Patient Data taskforce) and ideally participant co-development. They will include details as to: 1) the parties and data involved; 2) the way in which the data will flow and be combined; 3) the intended purpose for this data; 4) how we will minimise risks; 5) how we will respect existing objections and how to register any new objections.

The initial fair processing will likely differ from those typically used at a study-specific level: given the urgency of the situation and that social distancing restrictions may impact on studies ability to contact individuals using postal mailings or fieldwork visits. Where possible, the fair processing should still involve direct contact with participants in addition to social media and website mechanisms.

We accept that participants have actively objected to the study's use of their health records, this will be respected and applied to this use case (regardless of any permissive powers). Where individuals have set an Opt Out preference with their GP then this will also be respected where possible (i.e. where this status is known within the NHS data flows described above) unless the individual has also provided an explicit consent to the study.

9. Data Processing & Cleaning

We will take a relatively light touch approach to data processing and cleaning. The rationale being related to expediency and it is not possible to check individual values when making secondary use of records: therefore any processing decision is likely to be subjective and may not align with the needs of all resource users. Rather, we will rely on the source processing and checks (e.g. the NHS Hospital Episode Statistics records will have undergone iterative checks and cleaning as part of their centralisation, the LPS COVID-19 questionnaire will be cleaned at a study level).

The UK LLC will work with other initiatives in the NCS programme, the HDR UK Data & Connectivity NCS and other C-19/data science projects which are seeking to integrate complex health data across systems, across nations and across time. We will – where possible – adopt standardised approaches to shared data management challenges.

10. Discovery & Access

The UK LLC resource will be added to the HDRUK Gateway catalogue (https://healthdatagateway.org/) and will be documented in a 'Resource Profile' descriptive paper. The resource will be promoted through the contributing LPS to their research users, through data science networks (including HDRUK, www.hdruk.ac.uk), longitudinal study resources (such as the CLOSER consortium, www.closer.ac.uk) and funders.

The UK LLC application process will be based around the 'Five Safe's' governance framework and will have distributed and delegated review aspects (Figure 7). Applications will be coordinated through the HDR UK Gateway (www.healthdatagateway.org) which is managing all NCS data applications. The Gateway will provide discovery functionality (i.e. a metadata catalogue) and a standardised access request mechanism. The Gateway will also provide associated due diligence determining whether the applicant is a 'Safe Researcher' and at a high level whether the proposal constitutes "Safe Research' [https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf]. The UK LLC proposal review panel will receive the request from the Gateway and 'triage' it to ensure it is feasible and compatible with contracted requirements. The triage process will have a range of core functions:

- The UK LLC team will conduct project risk assessments (Data Privacy Impact Assessments), Data Protection compliance (e.g. maintaining data flow registers) and will ensure research transparency through adding the project details to a data use register;
- for *distributed approvals* where the authority to approve/reject applications is retained by the study, the UK LLC team will send relevant information out to each contributing study

(whose data will be used in the proposed project) and then record the decision to either include or exclude study data from the project specific dataset;

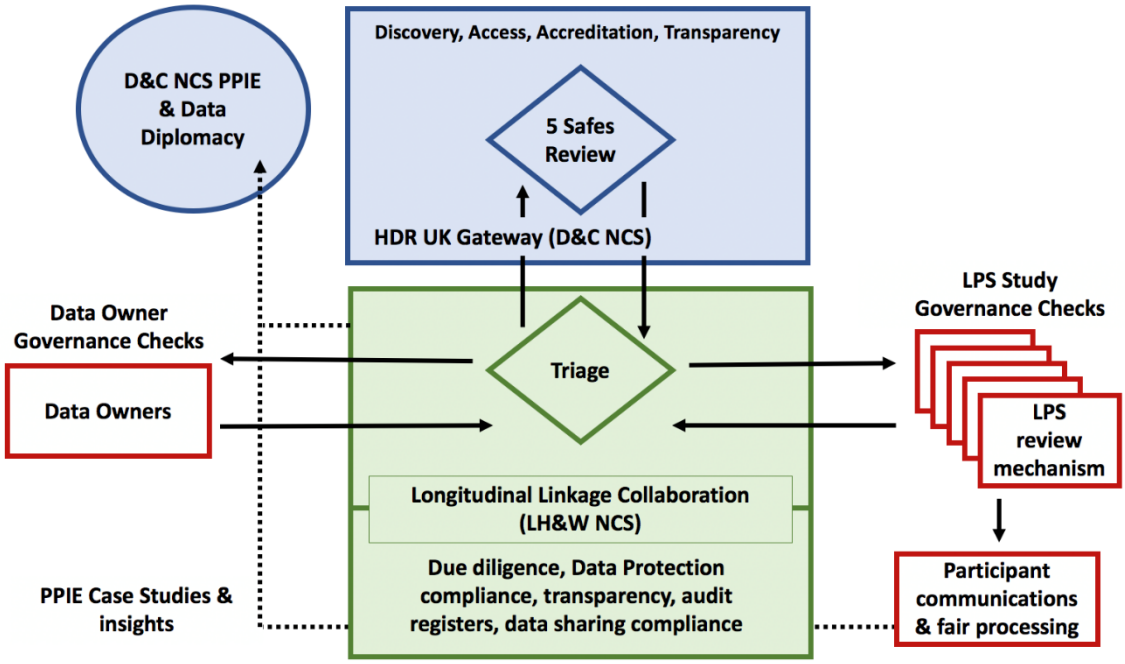
- for *delegated approvals* where authority to approve/reject applications is managed to a contracted framework by the UK LLC on behalf of contributing studies then the UK LLC will determine suitability against the agreed criteria and record the decision to either include or exclude study data from the project specific dataset.
- To assess project compliance with *data owner* conditions for enabling onward sharing of linked records[xiv]. Compliance decisions will be made against a framework agreed with the data owner and documented in a publicly available register. Where necessary, data use agreements and flows will be communicated back to the data owners (e.g. in annual reports).

The results of the application will be fed back into the HDR UK Gateway and from there to the applicant. The triage and distributed review process will be managed through Service Level Agreements in order to provide timely assessments.

The UK LLC proposal review panel will comprise UK LLC staff and representatives from contributing studies. We will seek to recruit lay public/patient members to the panel.

All project use will be made transparent through an LLC social media presence and a web-based register describing each data use in lay language. Examples of LLC outputs and research will be fed to contributing studies and the Public & Patient Involvement and Engagement (PPIE) activities being conducted through the HDRUK UK Data and Connectivity (D&C) NCS group as exemplars to demonstrate the value of the programme, to help ensure transparent data use and as a tool for PPIE (to provide use cases).

Figure 7: UK LLC data access approvals mechanism



11. UK LLC Output Reviews

In order to generate 'safe outputs' all outputs (with the exception of permitted identifiable data flows of study data back to the contributing study) leaving the UK LLC TRE will be reviewed for disclosure risk (within the boundary of the TRE). The output review process will be based on the framework for the SAIL databanks review process which is ISO27001 compliant and certified as meeting Digital Economy Act 2017 and UK Statistics Authority code of practice requirements. This framework considers statistical risks present in tables and other output forms (statistical covariate information, graphs and other data representations). Only anonymous aggregated information will be permitted to leave the TRE.

All users will be required to be Office for National Statistics Approved Researchers. The training needed to gain this accreditation includes awareness of output disclosure risk.

12. Descriptive analysis, assessments of representativeness and statistical weightings

We will process the data to conduct descriptive and documentary analysis. This work aims to inform research users and those considering research findings about the nature of the resource.

We will describe the aggregate UK LCC population in terms of the participants socio-economic, demographic and geographic distributions. We will seek to assess data quality, define the UK LLC denominator and to reference this clearly back to the contributing studies enrolled sample. We will seek to account for why individuals who are enrolled into a contributing study are not present in the UK LLC (e.g. dissent, linkage error). The outputs of this work are designed to inform data discovery

(e.g., through publishing 'resource profile' and 'data note' publications), to inform assessments of suitability for given research questions (e.g., to quantify sample sizes of population sub-groups), to inform users about the nature of the sample and its population distributions in order to inform analysis, and to help policy makers and other users to draw accurate inferences from research findings.

Some of the LLC contributing studies have population representative samples. In theory, this form of sampling approach enables the generation of accurate population inferences which can be generalised to the wider population. However, all longitudinal studies suffer from participant attrition and it is known that those continuing to participant tend to have different health and social characteristics to those lost to attrition. If uncontrolled, this can generate biased research findings and lead to situations where vulnerable sub-groups are excluded from the research and marginalised from the benefits of research (for example, groups whose health or social circumstances act as a barrier to continuing participation). To address monitor this, we will conduct statistical assessments of representation and use information from linked health and social records to develop statistical weightings to correct for the impact of attrition and changing population characteristics. Information from the linked records will also be used to inform statistical procedures (inverse probability weighting, multiple imputation) to address missing data in order to generate more accurate estimates (see for example, Cornish et al. for an illustration of the benefits of this in a longitudinal study context.[xv]

This programme of work will be designed to fit to international research reporting guidelines (e.g., the RECORD statement[xvi] for reporting on the use of linked routine records in applied findings, the GUILD statement[xvii] for reporting on linkage process) and will be aligned with data quality assessment criteria being developed by stakeholders such as the Office for National Statistics, and increasingly used within UK government.

13. The Longitudinal Health & Wellbeing National Core Study COVID-19 research programme

The Covid-19 pandemic and related government actions to supress viral transmission has resulted in a health crisis situation and will continue to have extreme health and social impacts worldwide for the mid to long-term. Our aim through the LH&W NCS is to quantify the health and social impacts of the pandemic, to understand underlying mechanisms, and to identify at risk groups. The contributing studies in the UK LLC have rich, often whole of life, antecedent data on health (both physical and mental) status and socio-demographic characteristics. Many have genotyping and dense biomarker characterisation. Participants of studies within this consortium have completed repeat questionnaires including items on C-19 infection diagnosis, testing and symptoms, health behaviours, physical and mental health, receipt of health care, employment, education, household composition and finances, and family and social relationships. Sample questionnaires are available [xviii,xix,xx,]. Further self-reported C-19 status data will be accessed through linkage to the 'Zoe' Covid-19 symptom tracking app[xxi]. This will provide information recorded at a high temporal frequency from early in the pandemic. Thus, these cohorts are well placed to understand not only the consequences of the spectrum of infection itself, but also the population effects of viral suppression measures.

Questionnaire outputs

These linked questionnaire data alone have already yielded important information. Health behaviours have changed, becoming more favourable in the affluent, and less favourable in disadvantaged groups. The proportion of people living in poverty has increased by 50%. Psychological distress has also increased, particularly in the young and in women[19]. In young adults, the number of people with anxiety has doubled. Those with pre-existing mental health conditions, and experiencing socioeconomic adversity are at especially high risk of poorer mental health during COVID-19[20,21] . These effects appear persistent, even with relaxed restrictions. Mental health service use has fallen[22,23] , suggesting people are delaying seeking help despite a worsening of mental health. However, it is unclear if these results represent a reaction to the pandemic which will subsequently reduce, or an early indication of a physical and mental health crisis which will continue as the pandemic unfolds and in the aftermath of the pandemic. Continued monitoring of health and health service use is essential to fully understand both the short and long-term impact of COVID-19.

Linkage with health and administrative records

Linked health records to these studies, the Longitudinal Linkage Project (LLP) will provide up-to-date information on Covid-19 testing, symptoms, treatment and outcomes. Linked data will be used in conjunction with the extant rich pre Covid-19 and questionnaire data. It will be used to inform statistical models and also to inform triangulation assessments of data quality and of potential bias in the contributing LPS (in order to provide supporting evidence to those seeking to understand findings emerging from the LH&W NCS programme).

These linked data will be of unique value in underpinning a programme of research on C-19 informed by the data assets within the UK LLC:

1) Understand patterns and predictors of infection, (including re-infection) and disease outcomes (such as 'long covid'), and the role of antecedent and current health behaviours, health status, medication use, sociodemographic status, built and natural environmental factors, in impacting these outcomes.

2) Explore population level changes to physical & mental health, including hospital admission and mortality, in association with viral suppression measures and how these relate to changes in health behaviours.

[19] Henderson, M., et al.(2020) Mental health during lockdown: evidence from four generations - Initial findings from the COVID-19 Survey in Five National Longitudinal Studies. London: UCL Centre for Longitudinal Studies.
[20] Bann D., et al. submitted (https://www.medrxiv.org/content/10.1101/2020.07.29.20164244v2)
[21] M Benzeval et al. (2020) Understanding Society COVID-19 Survey April Briefing Note: Health and Caring, Understanding Society Working Paper No 11/2020, ISER, University of Essex.
[22] Kwong et al., submitted (https://www.medrxiv.org/content/10.1101/2020.06.16.20133116v1)
[23] Niedzwiedz CL., et a (2020). Mental health and health behaviours before and during the initial phase of the COVID-19 lockdown: longitudinal analyses of the UK Household Longitudinal StudyJ Epidemiol Community Health. doi: 10.1136/jech-2020-215060

3) Investigate the role of socio-economic and neighbourhood/environmental factors in determining population level impacts to physical and mental health, to identify both groups at risk, and factors which offer resilience to adverse outcomes.

4) Analyse changes in health care service use (using self-reported data and NHS health records) to determine if patterns in these have changed during the C-19 pandemic. Linked NHS records will also inform consideration of service use in relation to pre-pandemic health status and regular service interactions (e.g. screening, health reviews and routine service take up (such as annual seasonal flu vaccinations)).

The LH&W NCS is also designed to be responsive to health and government policy makers. Through the LH&W 'Policy Exchange' function, senior scientists, health practitioners and planners and government policy makers will be able to set priority research questions (through the HDRUK and SAGE 'research funnel' and also directly to the LH&W NCS) and ask for insights to these from the longitudinal studies. This suggests a dynamic research programme – within the bounds of providing evidence to support the C-19 response – which accommodates changing needs reflecting the evolving pandemic and broader socio-economic context. It is envisaged that many requests for evidence will have short time frames and will take the form of rapid synthesis of available evidence; this will be coupled with in depth epidemiological and social science investigations using standard research methodologies (but ensuring quick flow of insights through publishing 'pre-prints' and through informing the HDRUK led briefings to the SAGE committee).

14. Sustainability

"One lesson that is very important to learn from this pandemic, and for emergencies in general, is that data flows and data systems are incredibly important. You need the information in order to be able to make the decisions. Therefore, for any emergency situation those data systems need to be in place up front to be able to give the information to make the analysis and make the decisions".

Sir Patrick Vallance[24]

The C-19 pandemic has emphasised the need for new ways of working within the research community: which enable the secure and sustainable co-location of longitudinal research data which is augmented with regularly updated flows of health and social records. This will be a research resource for UK emergency response: for the ongoing assessment of the impact of C-19 and other emergency situations such as the impacts of climate change, future pandemics and economic or social shocks. The objective is that the UK LLC provides a coordinating resource for linkage in LPS that is more responsive to contemporary circumstances and thus better suited to informing evidence-based policy decisions.

[24] Providing evidence to the Houses of Common's Science and Technology Committee on the UK Science, Research and Technology Capability and Influence in Global Disease Outbreaks. July 2020. Available from: https://committees.parliament.uk/oralevidence/701/html/

There is therefore a strong view that the UK LLC infrastructure established through the NCS should be a persistent resource of the UK population data science community. The primary funders of longitudinal research in the UK (the Medical Research Council, the Economic and Social Research Council and the Wellcome Trust) have convened 'Population Research UK' initiative to scope new ways of working across the LPS community. From late 2020 HDR UK will lead a PRUK scoping year which will consider how initiatives such as UK LLC suggest a new way of realising new scientific opportunities and reducing burden on data owners (e.g. the NHS) and maximising the cost efficiency of this process.

The UK LLC will play a full part in the PRUK scoping year. We will investigate options for funding sustainability (central research infrastructure support, cost recovery models) and conduct cost/benefit assessments for contributing LPS.

We will also consider the acceptability of the UK LCC persisting beyond C-19 with LPS participants and with the public through the PPIE initiatives within the HDR UK Data & Connectivity NCS.

15. Funding

This initiative is being supported by UK Government as a component of the Longitudinal Health & Wellbeing National Core Study. In turn, this is a theme within the National Core Studies for COVID-19 research initiative. The financing for this study is initially scheduled to run until 31.03.21 and has been provided by HM Treasury and is administered via the UK Medical Research Council on behalf of UK Research & Innovation.

16. References

[i] Medical Research Council. Strategic Review of the Largest UK Population Cohort Studies. Medical Research Council; London: 2014. www.mrc.ac.uk/populationcohortreview.

[ii] Northstone K, Haworth S, Smith D, Bowring C, Wells N, Timpson NJ. The Avon Longitudinal Study of Parents and Children-A resource for COVID-19 research: Questionnaire data capture April-May 2020. Wellcome Open Research. 2020 Jun 10;5(127):127.

Northstone K, Smith D, Bowring C, Wells N, Crawford M, Haworth S, Timpson NJ. The Avon Longitudinal Study of Parents and Children-A resource for COVID-19 research: Questionnaire data capture May-July 2020. Wellcome Open Research. 2020;5.

[iii] https://wellcomeopenresearch.org/articles/5-228

[iv] Kwong AS, Pearson RM, Adams MJ, Northstone K, Tilling K, Smith D, Fawns-Ritchie C, Bould H, Warne N, Zammit S, Gunnell DJ. Mental health during the COVID-19 pandemic in two longitudinal UK population cohorts. medRxiv. 2020 Jan 1.

[v] Burton J, Lynn P, Benzeval M. How Understanding Society: The UK household longitudinal study adapted to the COVID-19 pandemic. InSurvey Research Methods 2020 Jun 2 (Vol. 14, No. 2, pp. 235-239).

[vi] Carter, P, Laurie, G & Dixon-Woods, M 2015, 'The Social Licence for Research: Why care.data Ran Into
Trouble', Journal of Medical Ethics, vol. 41, no. 5, PMID: 25617016, pp. 404-409. https://doi.org/10.1136/medethics-2014-102374

[vii] Boyd A, Coleman G, Spence E, Park A, Hardy H. (2019). An outline framework for the efficient onward-sharing of linked Longitudinal Study and NHS Digital records. London, UK: CLOSER,

University College London. Available from: https://www.closer.ac.uk/wp-content/uploads/091219-NHS-Digital-and-LPS-onward-sharing-report.pdf
[viii] Trusted Research Environments (TRE): A strategy to build public trust and meet changing health data science needs. April 2020. Health Data Research UK. Available from: https://ukhealthdata.org/wp-content/uploads/2020/04/200430-TRE-Green-Paper-v1.pdf
[ix] https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf
[x] https://committees.parliament.uk/oralevidence/701/html/
[xi] Burton PR, Murtagh MJ, Boyd A, Williams JB, Dove ES, Wallace SE, Tasse AM, Little J, Chisholm RL, Gaye A, Hveem K. Data Safe Havens in health research and healthcare. Bioinformatics. 2015 Oct 15;31(20):3241-8.
[xii] https://ico.org.uk/media/1061/anonymisation-code.pdf
[xiii] Rodgers SE, Lyons RA, Dsilva R, Jones KH, Brooks CJ, Ford DV, John G, Verplancke JP. Residential Anonymous Linking Fields (RALFs): a novel information infrastructure to study the interaction between the environment and individuals' health. Journal of public health. 2009 Dec 1;31(4):582-8.
[xiv] Boyd A, Coleman G, Spence E, Park A, Hardy H. (2019). An outline framework for the efficient onward-sharing of linked Longitudinal Study and NHS Digital records. London, UK: CLOSER, University College London. Available from: https://www.closer.ac.uk/wp-content/uploads/091219-NHS-Digital-and-LPS-onward-sharing-report.pdf
[xv] Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. International journal of epidemiology. 2015 Jun 1;44(3):937-45.
[xvi] Benchimol EI, Smeeth L, Guttmann A, et al. The Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med. 2015;12(10):e1001885
[xvii] Ruth Gilbert, Rosemary Lafferty, Gareth Hagger-Johnson, Katie Harron, Li-Chun Zhang, Peter Smith, Chris Dibben, Harvey Goldstein, GUILD: GUidance for Information about Linking Data sets, Journal of Public Health, Volume 40, Issue 1, March 2018, Pages 191–198, https://doi.org/10.1093/pubmed/fdx037
[xviii] https://bristol.ac.uk/alspac/researchers/wellcome-covid-19/
[xix] https://www.understandingsociety.ac.uk/documentation/covid-19/questionnaires
[xx] https://cls.ucl.ac.uk/wp-content/uploads/2020/05/UCL-Centre-for-Longitudinal-Studies-COVID-19-Online-Survey-Questionnaire-Wave-1-April-2020.pdf
[xxi] https://covid.joinzoe.com/