A national health data research capability to support COVID-19 research questions

14 April 2020

SAGE needs COVID-19 research questions to be rapidly answered to guide national (and international) decision making. The four nations of the UK need an approach to enable rapid research at scale and utilising existing research assets, talents and expertise across academia, NHS, government, charities and industry.

This paper proposes a legally compliant approach to link data and provide access to secure analytical environments for researchers to answer rapid COVID-10 related research questions, across the four nations, using cardiovascular disease as an exemplar. This has been developed by NHS Digital (NHSD), Health Data Research UK, National Institute for Cardiovascular Outcomes Research (NICOR) and the British Heart Foundation (BHF) and is generalisable to other conditions. It will be subject to review after the COVID-19 emergency, or after six months, whichever comes first.

We are seeking from SAGE:

i) **endorsement for this approach to enable it to be operational over the next 1-2 weeks**
ii) **'pull' for the answers to key policy and research questions via this approach**
iii) **support to request that data custodians make their data rapidly accessible through this route**
iv) **encouragement for funders to facilitate the scale-up and operation of this approach**

Objectives and measures of success

In this unique environment, there is a pressing need to address both service delivery issues and research to improve understanding and treatment. This approach aims to support rapid research using health data into healthcare outcomes associated with COVID-19 and into potential interventions to reduce the severity of those outcomes. It aims to be scalable, to represent all four nations of the UK, to provide safe and trustworthy access, and to avoid distracting the operational activities of the NHS.

Rapid research using health data requires:

- Good research questions, whose answers will rapidly benefit current clinical care and the public's health
- Skills in the rapid design, analysis and interpretation of data-driven studies
- Large scale data, ideally from all four nations, to enable statistically well-powered studies (and to identify regional differences)

- Pseudonymised datasets linked at an individual level to analyse patterns and test hypotheses[1]
- A computing environment with appropriate tools for data management and analysis, within which researchers can address high priority questions effectively and efficiently
- Ability, through open standards and open APIS, to link with international efforts of open research and data sharing, that harness data science to combat COVID-19
- An access and authentication process, supported by logging, that allows large numbers[2] of approved researchers to access data safely (safe data, safe projects, safe people, safe settings, safe outputs)[3]
- Furthermore, the information governance will need to be in place to ensure transparency and fair processing that complies with legal gateways, duty of confidence and current best practice (including the application of opt-out where required), research ethics and data minimisation.

The success of this approach will be judged by the speed and quality of the answers to priority research questions. The scalability of the approach in terms of number of researchers and research questions that can be concurrently addressed, along with public perception of trustworthiness will be additional measures of success.

Cardiovascular disease example

People with cardiovascular disease are one of the groups most likely to be directly and indirectly adversely affected by COVID-19. For example, data from hospitals across England show that the number of people seen in hospital with a suspected heart attack has halved since the beginning of March, raising concerns that people may be at greater risk of suffering long term heart damage, needing intensive care, or even dying as a result.[4] In the light of this, the clinical cardiovascular community, through Professional Societies and NICOR, proposed collaboration with NHSD/X and PHE to create a linked dataset without IG barriers.

In the last few weeks, highly effective collaboration has been established with a first imperative to focus on Acute Coronary Syndromes based on SUS+/HES, PDS/ONS, and NICOR datasets to describe presentation and deaths. The British Heart Foundation have brought together the BHF Data Science Centre, HDR UK, other academic stakeholders, to develop further the key research questions and to ensure the best access to these combined NHS datasets for research. These research questions are provided in Appendix 1.

---

[1] Although this data will be pseudonymised, it will remain identifiable and sensitive. As such it will need information governance appropriate to identifiable data.

[2] 100s to 1000s

[3] Aligned with UK Health Data Research Alliance Principles for Participation for data custodians

[4] https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2020/april/drop-in-heart-attack-patients-amidst-coronavirus-outbreak

Proposed approach

The proposed approach is condition agnostic. It uses cardiovascular disease as an exemplar condition, but it is fully generalisable to other data sets and conditions such as cancer, diabetes and mental health.

There are three parts to the approach, A. Research Question Funnel, B. Linked data via Trusted Research Environments, C. Information Governance and Access.

A. Research Question Funnel

The best research questions might come from clinicians observing what's happening at the front line, life science companies working on treatments for similar viruses, epidemiologists observing implications for patients with cardiovascular disease, or others. We have created a Darwinian question funnel that helps the 'fittest' questions to be answered fastest:

- Priority being given to those commissioned by CSA/CMO, and through SAGE
- Supported by a simple, open route for anyone to pose research questions HDR UK COVID-19 Knowledge + Skills Matchmaker
- A weekly, objective and transparent prioritisation process to identify which research questions need to be answered quickest: the NIHR/HDR UK prioritisation process with panel members drawn from across the four nations.

B. Linked NHS data in English, Scottish, Welsh & Northern Irish Trusted Research Environments

By linking de-identified data together within each jurisdiction and by making it accessible in trusted research environments, many questions can be answered by different researchers, using the same datasets. Note the proposal is not to link datasets across jurisdictions.

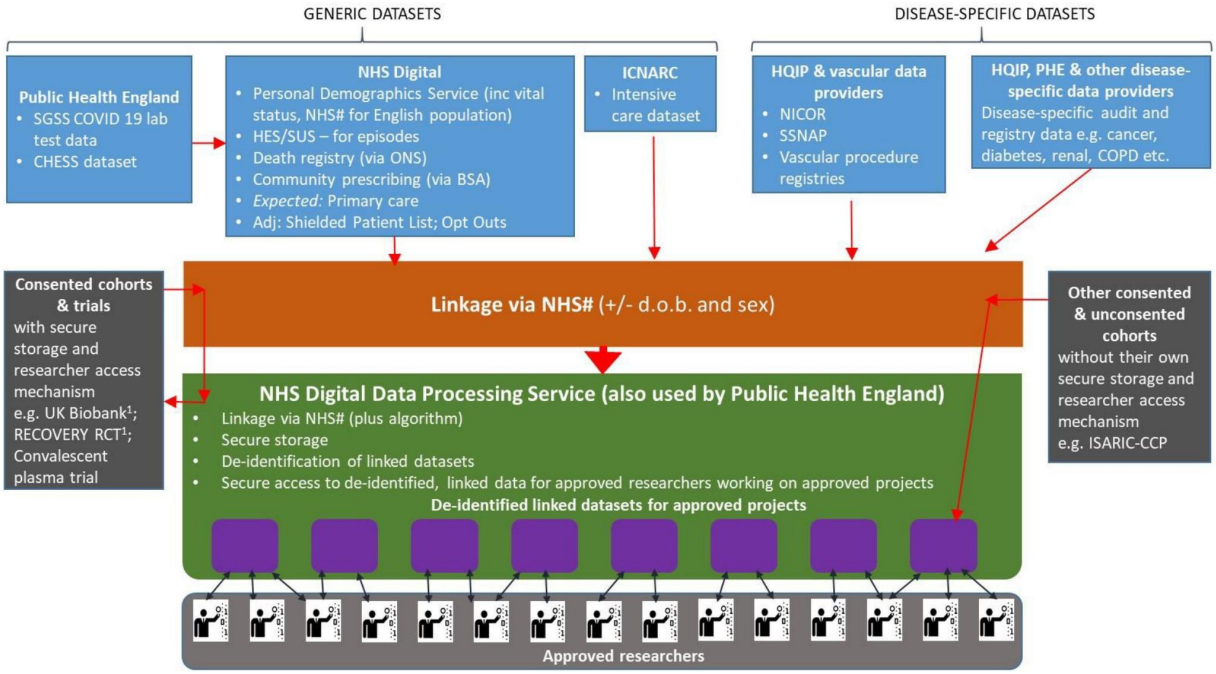
To respond rapidly we propose utilising existing trusted research environments, which have established information governance processes. These are:

Location*	Trusted Third Party	Trusted Research Environment
England	NHSD	Data Processing Services (DPS)
Wales & NI	NHS Wales Informatics Service (NWIS)	Secure Anonymised Information Linkage (SAIL) databank
Scotland	Edinburgh Parallel Computing Centre (EPCC) on behalf of Public Health Scotland	Scottish National Data Safe Haven
Northern Ireland	Honest Broker Service	Business Services Organisation (BSO)

**This is where the TRE is located; it does not mean that the TRE can only host data from that geography*

These Trusted Research Environments would host and link datasets (within their jurisdiction) and provide access to de-identified linked data to approved researchers. The following figure sets out the generic datasets and the disease specific datasets in the England TRE:

Figure 1. Linkage and secure access for approved research to nationally-collated, generic and disease-specific datasets (cardiovascular and other) in England[5]



Similar arrangements developed for linkage of key datasets in Scotland, Wales & Northern Ireland are described in Appendix 2.

This approach is open to other trusted research environments, and we would expect others to participate once this approach is established, for example 'enhanced cohorts' with genomic and other biological data with particular requirements that may not be met by the national TREs.

---

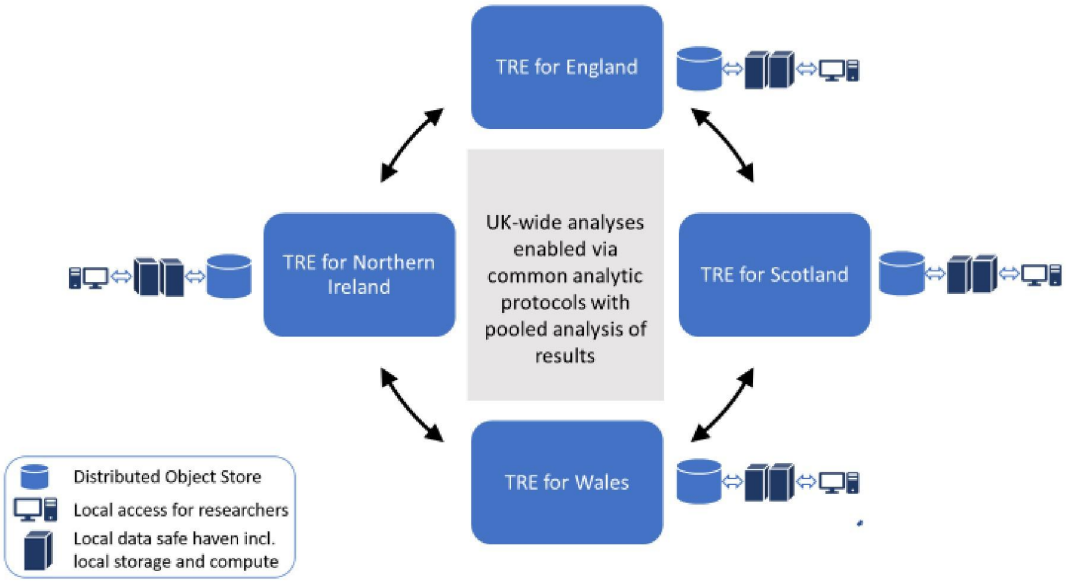
[5] SGSS: Second Generation Surveillance System; CHESS: COVID-19 Hospitalisation in England Surveillance System (CHESS); HES: Hospital Episode Statistics; SUS: Secondary Uses Service; BSA: Business Services Authority; PDS: Personal Demographic Service; ICNARC: Intensive Care National Audit and Research 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

NB: UK Biobank is currently set up to receive data feeds of these datasets separately from PHE, NHSD and ICNARC, and receives linked primary care data for participants in England direct from the primary care computer system suppliers (TPP and EMIS). RECOVERY (and others) are currently pursuing a similar route for linkage to data from PHE, NHSD and ICNARC

Linkage within jurisdictions would be conducted by the trusted third party, with linked data for the whole population in that jurisdiction hosted in the TRE. Linkage would be achieved using the NHS (or CHI in Scotland, or Health and Care number in NI) number.

UK-wide analyses would be achieved through a federated network of TRE's across the four nations of the UK, by sharing of analytical outputs, as shown below:

Figure 2. Federated Network of Trusted Research Environments



C. Information Governance and access

The discovery, access, authentication and analysis process has the following components:

- **Understanding of datasets:** The Innovation Gateway, will be the open source way of researchers understanding the availability of datasets and how to access them.
- **Prioritisation:** Research questions that have been prioritised through the Research Question Funnel will be prioritised for rapid access.
- **Application:** Each TRE and Trusted Third Party will provide a clear data access process, capable of operating at the scale and pace required for rapid response to prioritised research questions. Each request will need to be clear on its legal basis and purpose, and how it meets the policy and legal framework for each country. For example, within GDPR, COPI notices, Common Law on dissemination, and for England, whether National Data Opt-Out applies. Each will be required to have the relevant Transparency Notices.

- **Access:** Where access granted, each of the TREs will provide remote access for accredited academic, NHS, and government researchers to support multi-disciplinary team science. Each TRE will provide a practical approach for researcher accreditation and make this transparent on the Gateway so that the requirements are clear. The access journey for prioritised research questions will be tracked and transparent.
- **User Management:** Expedited Trusted access management processes will be used to assess and authorise remote access to data in each TRE.
- **Audit trails:** for all queries and activities and the resources to analyse and publish on the use of the data
- **Public transparency:** including publication of data access requests granted, stating the datasets, the research question, the legal purpose and the name of the organisation. In addition, patient groups included; and how subject access requests and other data subject rights, opt-out and ethics will be handled.
- **Data minimisation**: ensuing the fields and duration of data is minimised

The proposal will require further detailed development for each of these Information Governance and access arrangements.

Funding

Utilising existing assets and expertise, enables a fast and efficient approach. Additional funding is required to enhance and accelerate:

- The scale of the Access Management Teams in each TRE to enable speed of response
- Establishing and maintaining new data flows and linkage with frequent updates (e.g., fortnightly)
- Platform processing capacity
- Software to accommodate specific analytical requirements
- A small team to operate the question funnel and end to end performance metrics

In addition, the researcher costs will need to be funded by re-purposing existing grants or through new funding sources.

Roles & Responsibilities

All **data custodians** who are members of the **UK Health Data Research Alliance**, supported by HDR UK, will be asked to make their data accessible for linkage to support rapid response to COVID-19 research questions through each national proposed route.

Where data is required to answer the prioritised questions, but custodians are not yet members of the Alliance, these custodians will be asked to join the Alliance and in parallel will be asked to make their data accessible through this route.

The UK Health Data Research Alliance network of **patient and public advisory groups** will be asked to involve patient and public representatives in the design of this approach and the access decision making processes to ensure they are trustworthy.

NHSD, SAIL, the Scottish National Data Safe Haven and the BSO Honest Broker Service will be responsible for ingesting data flows, data linkage, operating the TRE in line with the Five Safes model, and managing the access management process in a secure and rapid way to respond to the priority questions.

NICOR will continue to support NHSD, to deliver to SAGE in an expeditated fashion the key cardiovascular indicators relevant to service delivery on an ongoing basis. Together with the professional societies and relevant clinical colleagues, it will collect, curate, analyse and make available the data to all relevant stakeholders through the proposed research platform. NICOR will work with HDR UK, NHSD to ensure that the process is compliant with the information Governance framework adopted.

The **BHF Data Science Centre**, in partnership with NICOR and the NIHR-BHF CV Partnership (which brings together the combined BHF and NIHR-supported cardiovascular research community), will provide the expertise in priority, data-driven, cardiovascular questions, as well as in study design, analysis and interpretation to ensure good quality research questions are being fed into the question funnel, and that they are appropriately resourced with the appropriate, accredited, analytical expertise to answer the questions once they have access to the data and environment.

HDR UK will manage the research question funnel and operate the end to end performance tracking information and provide summary updates to SAGE, Trusted Third Parties, data custodians and PIs for the research questions so that any delays and obstacles are visible and quickly addressed.

Expected scale up to other conditions

The approaches to link multiple sources of clinical data within a trusted environment and the research approach that we propose are generic, and will not only meet the immediate cardiovascular ask, but will also enable rapid answers to a wide range of questions without disease-specific dataset linkages. For example, vaccination research questions, such as whether the MMR vaccine might provide protection against COVID-19 can be tested with the linked primary and secondary care data.

New disease-specific data to link into the environment will be prioritised based on the importance of the questions, whether a disease-specific dataset(s) exists, and what its added value is alongside the generic datasets. HDR UK will facilitate a datasets for linkage prioritisation exercise to help focus linkage efforts.

Indicative timescales

By

- 21 April 2020:
  - For the cardiovascular example presented here we would have the initial generic and disease specific datasets accessible and access granted to first research applicants for each of the Trusted Research Environments.
  - First set of key questions relevant to SAGE CV COVID policy for the short, medium and long term available to drive next data sets to be ingested
- 28 April 2020 (and weekly thereafter)
  - Summary of prioritised questions
  - Report on speed and scale of access
  - Summary of emerging research findings

Relationships with other health data initiatives

- **NHSX Palantir-** This approach would be complementary and supportive of front line activities developed within the NHS, for example the NHSX COVID-19 datastore that will provide national organisations responsible for coordinating the response with secure, reliable and timely data to support decision-making[6]. The processes will be closely aligned, in particular access management and Information Governance.
- **Innovation Gateway** - will make metadata for datasets involved in this approach, discoverable and clearly signpost the access approach so that researchers understand what data is available and how to access it
- **Clinical Trials** – for example, NHS DigiTrials – is running in parallel, and using the same source datasets in NHS Digital, to support e.g. the RECOVERY trial on follow-up and outcome; and to provide a cohort for NHS Blood and Transplant to contact donors into the Convalescent Blood Plasma trial.
- **Health Data Research Hubs** – the other hubs, including DATA-CAN, Discover NOW, BREATHE, Pioneer (including DeCOVID), Gut Reaction will support the roll-out of this approach to other conditions (akin to the role the BHF Data Science Centre is undertaking in the cardiovascular example and NHS DigiTrials for clinical trials)
- **Genome sequence alliance (COG-UK) -** Sequencing of viral genomes, of those hospitalised, and also NHS workers to create family trees of virus genotypes, and different viral types, linking records of individuals with viral types-> understand specific genicity. Using the same source data from PHE and NHS Digital (in England), and using the MRC Cloud Infrastructure for Microbial Bioinformatics (www.climb.ac.uk)
- **Health Data Research UK Multi-omics Consortium –** This is bringing together about a dozen cohorts, collectively comprising some 400,000 participants. Taken together with UK Biobank, the cohorts provide a platform of ~1 million well-characterised research participants to study COVID-19

[6] https://healthtech.blog.gov.uk/2020/03/28/the-power-of-data-in-a-pandemic/

and CVDs (and other conditions), as the pandemic unfolds in the UK. The proposed approach in this paper would provide an environment to include analysis of these datasets.

- **NIHR-BHF Cardiovascular Partnership -** This existing Partnership which brings together the combined BHF (Research Excellence and Accelerator Award Centres, BHF Chairs) and NIHR (CV themes within BRCs) cardiovascular research infrastructure has created an operational framework to implement and resource coordinated strategy to (i) identify and prioritize Covid-19 research questions that respond to the Government's need, maximize impact and patient benefit and avoid duplication (ii) garner the full potential of the UK cardiovascular clinical and research community in a rapid, collaborative and transparent fashion (iii) align with other national and international initiatives. The Partnership will consider applications weekly and those that are approved and are data-driven will be channeled through the BHF Data Science Centre under this proposal.

Risks of this approach and proposed mitigation

The following risks have been considered in the development of this approach and suggested mitigation:

Risk	Mitigation
TREs are unable to support type of analysis that is needed (e.g. each TRE will be limited by the existing data within it, or that can be linked within the CV-19 time period).	Ensure each is supported in meeting a common set of user needs/stories across each nation. Each TRE data access environment has been designed to support multi-tenancy. Provide additional toolsets onto each TRE, as required (e.g. the NHS Digital Data Processing Service supports ACL, Python and SQL; and has set up a separate PHE environment to support R. This can be done for other CV19 projects).
Data access requests processes are not responsive enough for current emergency	Focus on the outputs of the prioritisation process, to ensure each request is addressed appropriately. Additional resource to support each TREs IG access process, while balancing the need to retain public trust. HDRUK to coordinate between each TRE.
Confusion over data request routes with NHSX Data Store	Make the distinction that the NHSX is focused on direct care and prediction/forecasting of direct care
Public / patient privacy concerns undermine proposal	Each Trusted third party to lead the transparency and fair processing of the data in each environment with explicit leadership from their respective public and patient advisory groups. Provide transparent communications on the proposed approach, responsibilities and a route for the public to raise concerns.

	Transparent and clear articulation of the 'five safes' for each TRE. For example NHS Digital has safe data (Privacy Enhancing Technologies to de-identify, use of derivations and restrictions), safe projects (only those with IG approval), safe people (staff with DBS check and above), safe settings (encrypted and cyber-tested platform), safe outputs (related to scope of project; researcher responsible for statistical Disclosure Control)
Legal basis for accessing some data ceases with COVID-19 emergency powers ceasing (or being extended) in Sep 2020, leading to time limited impact of research database.	DHSC/NHSX carefully consider the requirements for the research database at point of cease/extension
Linkage capability is insufficient to meet research requirements	In England, capability exists today via the Master Patient Service (MPS) component of Data Processing Service. MPS is an authoritative secondary use list of patients in England, fed daily from the direct care Personal Demographics Service (56m population).

Authors:

Caroline Cake, Health Data Research UK

Tom Denwood, NHS Digital

Cathie Sudlow, BHF Data Science Centre

David Seymour, UK Health Data Research Alliance

Nilesh Samani, British Heart Foundation

John Deanfield, NICOR

14 April 2020

Appendix 1: Cardiovascular Disease Research priorities to inform clinical and public health policy response to COVID-19

A. Key questions of relevance for UK cardiovascular disease data science community in the COVID 19 pandemic

(1) What is the impact of pre-existing cardiovascular disease,its risk factors (e.g. hypertension, diabetes, smoking. Ethnicity, gender) and cardiovascular medication on outcome in COVID 19 infection?

(2) What are the cardiovascular complications of COVID 19 infection?

(3) What is the impact of the NHS, public health and population response to the COVID 19 crisis on non-COVID diseases, in particular on the presentation, management and outcomes of other cardiovascular diseases?

(4) What are the molecular/multi-omic determinants and consequences of COVID 19 infection and what is the role and mechanism of cardiovascular disease in susceptibility to infection?

(5) What specific interventions might improve outcomes from COVID 19 infection in patients with pre-existing cardiovascular disease or cardiovascular complications of COVID 19?

B. Data-enabled routes to address these questions

(1) How does cardiovascular disease (and its treatment) affect outcome in COVID 19 infection?

This requires the study of cohort(s) of people with COVID 19 disease, characterised according to their cardiovascular disease history, medication (e.g. ACE inhibitors) and current status, with information on potential confounders (sex, age, socioeconomic status, health behaviours, other co-morbidities, drug treatments etc) and with follow-up information on their outcomes, both short term (e.g. within 30 days of admission to hospital, - death, requiring admission to ITU, requiring mechanical ventilation) and long term (e.g. death by cause at 6 months, one year and beyond post admission to hospital; recurrent stroke, MI and revascularisation events at 6 months, one year and beyond post admission to hospital)

Potential datasets that could address this question:

- *ISARIC CCP* (ISARIC CCP, as of 7 April 2020, n=9,000, increasing daily) – UK-wide NIHR-prioritised clinical characterisation of patients hospitalised with lab-proven or suspected COVID 19. Approval has been obtained for recruitment without consent.

  Substantially enhanced by linkages to primary care, secondary care (HES or devolved nation equivalents), mortality, ITU (ICNARC or SICSAG in Scotland) +/- community and hospital prescribing data.

  Additional linkages to NICOR (cardiac audits), SSNAP (stroke audit) and vascular registry (vascular surgical procedures audit) data would increase depth of cardiovascular disease characterisation.

- *ISARIC CCP with additional cardiovascular characterisation data (CAPACITY COVID, CAPACITY)*

- *All COVID 19 test +ve patients* from PHE (or devolved nation equivalent) lab testing data (as of 13 April 2020, n=69,329 in England, 6,067 in Scotland, 5,610 in Wales, 1,882 in N Ireland) linked to primary care, secondary care (HES or devolved nation equivalent), death registration, ITU (ICNARC/SICSAG) +/- community and hospital prescribing data.
- Additional linkages to NICOR, SSNAP and vascular registry data would increase depth of cardiovascular disease characterisation. (NB in theory all COVID 19 test +ve patients should be included in the ISARIC cohort but this is not yet the case.)
- *UK Biobank / other large multi-omic cohorts* linked to COVID 19 lab testing data, primary care, secondary care (HES), death registration, ITU (ICNARC) +/- community and hospital prescribing data. Additional linkage to NICOR, SSNAP and vascular registry data would increase depth of cardiovascular disease characterisation.

(2) What are the cardiovascular complications of COVID 19 infection?

This requires the study of cohorts of individuals with and without COVID 19 infection and/or individuals with different severities of COVID 19 disease with information on or follow-up for cardiovascular disease outcomes, both 'classical' CV outcomes (e.g. MI, stroke etc) and rarer, specific CV outcomes that may be a direct consequence/complication of COVID 19 infection (e.g. acute cardiac injury; myocarditis).

Potential datasets that could address this question:

- *ISARIC CCP* with linkages as above
- *ISARIC CCP with additional cardiovascular characterisation data (CAPACITY COVID) with linkages as above*
- *HIC-Cardiovascular COVID 19 initiative* (https://hic.nihr.ac.uk/)
- *Population wide linked data from England, Wales, Scotland and N Ireland:* COVID 19 lab testing data (to identify COVID 19 +ves) linked to primary care, secondary care (HES or devolved nation equivalents), death registration, ITU (ICNARC/SICSAG) +/- community and hospital prescribing data. Additional linkage to NICOR, SSNAP and vascular registry data would increase depth of cardiovascular disease characterisation.
- *UK Biobank / other large multi-omic cohorts* with linkages as above

(3) What is the impact of the NHS, public health and population response to the COVID 19 crisis on non-COVID diseases, in particular on the presentation, management and outcomes of cardiovascular diseases such as acute myocardial infarction and stroke?

During the current pandemic, people will continue to develop acute cardiovascular conditions (e.g., heart attacks, strokes, acute disturbance of cardiac rhythm, acute exacerbations of heart failure etc). Yet, emerging evidence from the UK and other countries is that hospital admissions of patients with acute

coronary syndromes have declined substantially in recent weeks, while patients are seeking medical help much later (or not at all) (BHF on COVID 19, ESC on COVID 19).

Datasets that could inform on time trends analysis to assess population-level monthly incidence and cause specific mortality before during and after the epidemic of ischaemic heart disease, stroke, heart failure, cardiac surgery, cardiovascular interventional procedures, vascular surgery etc… include:

- *Hospital admissions with diagnostic and OPCS procedural coding (HES and devolved nation equivalents)*
- *NICOR datasets*
- *SSNAP*
- *Vascular registry datasets*
- *Mortality data from death registries*
- *Primary care data*

These would not need to be linked to enable informative analyses (although linkage would enable a greater range of informative analyses).

(4) What are the molecular/multi-omic determinants and consequences of COVID 19 infection and what is the role and mechanism of cardiovascular disease in susceptibility to infection?

Potential datasets that could address these questions include:

- *UK Biobank / other large multi-omic cohorts* with linkages as above

(6) What specific interventions might improve outcomes from COVID 19 infection in patients with pre-existing cardiovascular disease or cardiovascular complications of COVID 19?

Addressing this question would require streamlined data-enabled RCTs of promising interventions.

Appendix 2: Linkage and secure access for approved research to nationally-collated datasets in Scotland, Wales and Northern Ireland

Figure 3: Linkage and secure access in Scotland

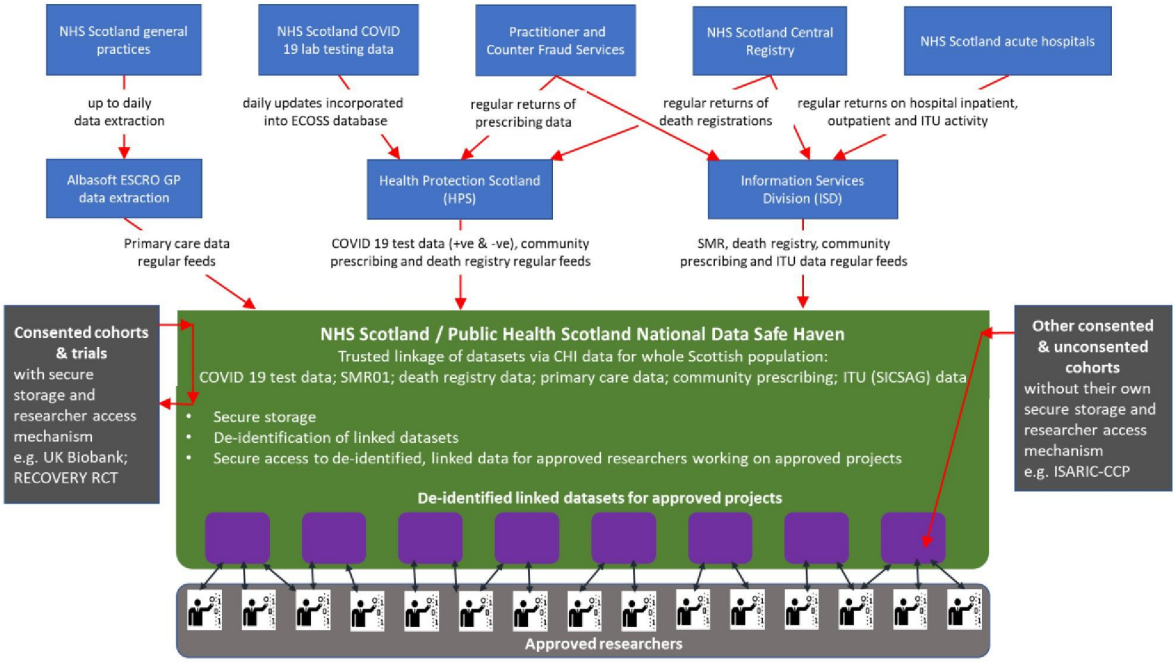
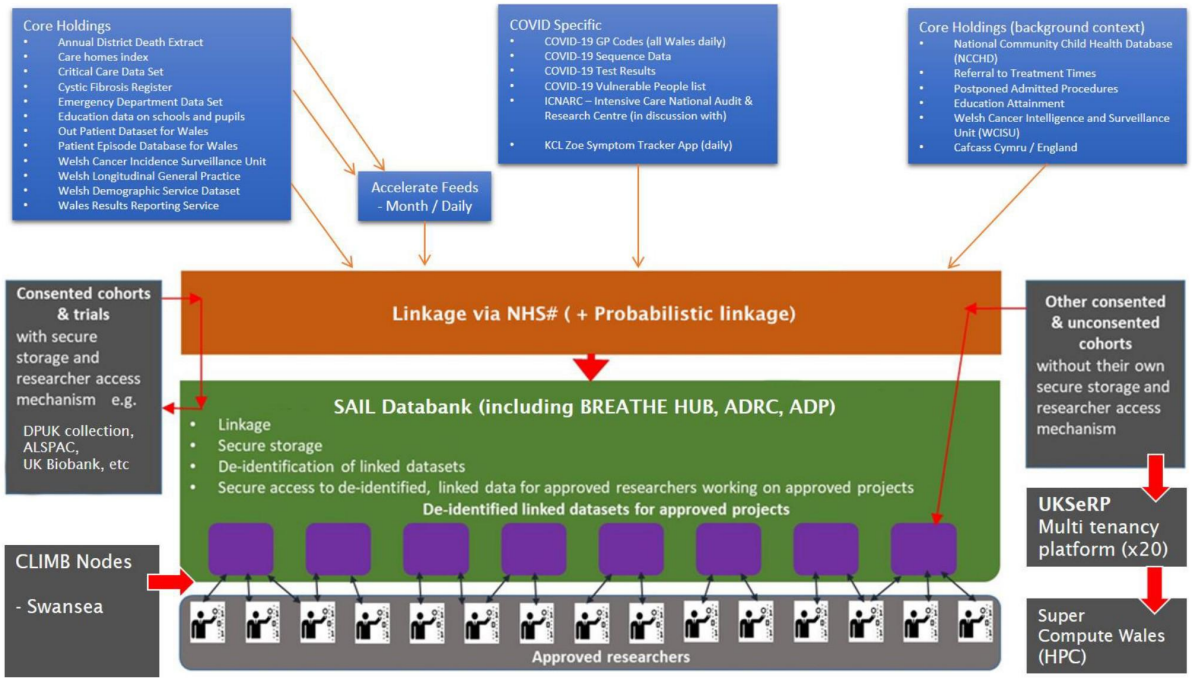


Figure 4: Linkage and secure access in Wales



UKSeRP = UK Secure eResearch Platform; ADRC = Administrative Data Research Centre; ADP = MQ Adolescent Data Platform

Northern Ireland: the Honest Broker Service - http://www.hscbusiness.hscni.net/services/2454.htm