Outbreak Data Analysis Platform:

The Manual

## WORK IN PROGRESS!

Contents

[WORK IN PROGRESS! 1](#_Toc107302468)

[Chapter 2 What is the Outbreak Data Analysis Platform (ODAP\*)? 3](#_Toc107302469)

[The Platform 3](#_Toc107302470)

[What is the purpose? 3](#_Toc107302471)

[What are the different environments/areas? 3](#_Toc107302472)

[Chapter 3 How is ODAP governed and managed? 5](#_Toc107302473)

[The Partnership 5](#_Toc107302474)

[The ODAP Delivery Team 5](#_Toc107302475)

[SPECIFIC ROLES 6](#_Toc107302476)

[Chapter 4 What is within scope of the ODAP? 7](#_Toc107302477)

[Chapter 5 What are the datasets are available for research analyses? What process(es) must I follow? What approvals must I obtain? 8](#_Toc107302478)

[Chapter 6 What is the ISARIC SPINE? 9](#_Toc107302479)

[Chapter 7 How do I add my dataset (within scope) into ODAP? 11](#_Toc107302480)

[Chapter 8 How do I access ODAP for data analysis? 12](#_Toc107302481)

[Data flows/Assumptions 12](#_Toc107302482)

[Single Data Controller Model 12](#_Toc107302483)

[Example text for COG-UK viral sequencing data within the UKHSA Data Controller agreement 13](#_Toc107302484)

[Chapter 9 BACKGROUND 14](#_Toc107302485)

[Chapter 10 Approvals 14](#_Toc107302486)

[Chapter 11 DATA CREATION PROCESS 15](#_Toc107302487)

[Chapter 12 Publication 16](#_Toc107302488)

[Chapter 13 How will I know when data I require for my analysis is available? And where it can be found? 17](#_Toc107302489)

[Chapter 14 How can I transfer my data files…? (not outputs but for further analyses) 18](#_Toc107302490)

[Chapter 15 How can I get outputs from ODAP? 19](#_Toc107302491)

[Chapter 16 How is ODAP communicated? Where can I refer people to about ODAP? 20](#_Toc107302492)

# What is the Outbreak Data Analysis Platform (ODAP)?

The ODAP is the overarching term for a range of computers supporting research within a specific scope. These include the ODAP TRE (formerly: flexible compute space; ultra; ultra2), and various project spaces in the National Safe Haven.

It is divided into two sections:

1. The “UNDER EMBARGO” section – data are wholly controlled by contributors
2. The “NO EMBARGO” section – data are accessible to appropriately-qualified researchers studying questions within scope

The foundation for the ODAP is the ISARIC Clinical Characterisation Protocol (CCP), an ethically-approved research study in the UK.

## The Platform

The Outbreak Data Analysis Platform (ODAP) is a data platform for (primarily/currently) Covid 19 research data, which include a combination of linked, curated data from UK sovereign data assets including the complete data resources of the ISARIC4C/CO-CIN, GenOMICC, PHOSP and UK-CIC studies, together with viral sequence data from COG-UK, and linkage to NHS clinical records (e.g. NHS England/NHS Digital, and Public Health Scotland) and structured clinical audit data. It also consists of a flexible high performance compute space/environment for data analyses.

## What is the purpose?

The purpose of the outbreak data analysis platform is to provide an accessible, usable data resource to enable research relevant to current (COVID-19) and future outbreaks. It aims to accelerate scientific understanding of new outbreaks for the benefit of patients and the protection of the public, through collaboration.

It will create a UK-wide capability by curating and linking outbreak relevant data from clinical records, research studies and audit data. It brings together key initiatives and leadership across the UK including ISARIC, COG-UK, MRC CLIMB and GenOMICC.

The platform combines a national Trusted Research Environment (TRE) infrastructure collocated with >£100M of world-class computational and data science capacity including the UK National Supercomputer, with a UK-wide governance framework.

Source: <http://odap.ac.uk/>

## What are the different environments/areas?

There are two areas on the platform: PHS Safe Haven and Flexible Compute Space (FCS). Both of these are Trusted Research Environments (TREs).

* **Safe Haven** – clinical data linking with various datasets
* **FCS** – ULTRA/ULTRA2 analyses data bringing together analyses/results from various different studies

Within these 2 areas, there is a further distinction of Embargo/Non-embargoed area. Embargo area is data not yet published or available for use, non-embargoed is data that can be accessed by bona fide researchers with no DSA required.

Source: WO

* In the safe haven environment there are three areas:
  + Restricted area: This contains sensitive data, e.g., data under a time embargo, an active trial and/or identifiable patient data. We will not be granting researchers access to this data.
  + Embargoed data: This contains pseudonymised patient level data curated and formatted to be used for research. Access permitted with a data sharing agreement.
  + Published data: This is the outputs from research and will contain summary/aggregate non-identifiable data only. This will be available in an open access format and available for research purposes without a data sharing agreement.
* There will be a secure API in place from the Flexible Compute Space to External TREs and UK Public Health Agencies. This is to enable data linkage across multiple data sets using unique identifiers to facilitate secure transfer if specified data fields e.g., viral sequencing.

Source: EM

Moni, Question: Is the API a suitable mechanism for transferring data from another TRE to the ODAP TRE? (Related to a question raised by Annemarie out of meeting)  
Rob, Andrew, Kenny: No, the preferred way to move data between another TRE and ODAP are the already established methods for that TRE.  
  
    Kenny, clarification: The future purpose of the API is more in line with providing a mechanism for quick pandemic updates to national health services.

Architecture doc – not for public consumption? Check with Rob.

# How is ODAP governed and managed?

TAKE FROM HDR UK DOCS

## The Partnership

Partnership: to strengthen the UK-wide NHS/academic partnership between the CLIMB COVID viral sequence analysis system, the Outbreak Data Analysis Platform, ISARIC-4C, COG-UK, the four UK Public Health Agencies, ICNARC, ONS and HDR UK. ​

The ODAP Partnership aims to create a new nationwide partnership to facilitate vital research on the impact of SARS-CoV-2 genetic and phenotypic variation on disease severity and vaccine efficacy by creating additional data linkages, automating data flows and analyses, and democratising access to research datasets with full agreement from the contributing studies.

Partnership: to strengthen the UK-wide NHS/academic partnership between the CLIMB COVID viral sequence analysis system, the Outbreak Data Analysis Platform, ISARIC-4C, COG-UK, the four UK Public Health Agencies, ICNARC, ONS and HDR UK.

The ODAP Partnership will enable new scientific analyses that are currently not possible at population scale:

The role of SARS-CoV-2 variation (>1 million viral genomes) in COVID-19 severity (ISARIC 4C / ICNARC) and outcomes (PHOSP).

Systematic analysis of the impact of vaccines and treatments (antivirals, etc), and identification of escape/resistance mutations.

Further linkage of UK research (e.g., immunology studies) and trial data (e.g., REMAP-CAP).

Develop automated processes to conduct analysis in near real-time to provide rapid multiparameter determination of new variants.

Continuing wider linkage of diverse datasets from individual studies that generate information on pathogen genomics, host genomics, surveillance, and patient outcomes are vital to build predictive models and systems to improve patient treatments and outcomes.

The Steering Group

ODAP steering group to provide advice, support and challenge to ODAP delivery team and will have voting rights on decisions not made by consensus.

## The ODAP Delivery Team

ODAP Delivery team to report progress to Data and Connectivity Delivery Group

## SPECIFIC ROLES

University of Edinburgh (as the lead institution for ODAP) is accountable to Health Data Research UK for the delivery of ODAP

HDRUK is accountable to UKRI for delivery of the Data and Connectivity National Core Study

Source: HDR UK slides (various)

# What is within scope of the ODAP?

The purpose of the ODAP is to facilitate biomedical research to advance understanding of severe infectious disease\* and other exposures of public health interest.\*\* Research within the ODAP is strictly limited to this purpose.

\* Severe infectious disease - this term describes all severe infectious agents, including new, re-emerging or therapy-resistant forms of existing infectious agents.

\*\* Other exposures of public health interest: this term describes new or unexplained poisoning, or exposure to harmful energy sources such as electromagnetic radiation.

# What are the datasets are available for research analyses? What process(es) must I follow? What approvals must I obtain?

A live [ODAP datasets mapping.xlsx](https://uoe.sharepoint.com/:x:/r/sites/ISARIC4C/DataInfrastructureAndGovernance/ODAP/_documents/ODAP%20datasets%20mapping.xlsx?d=w1dde5dbd6a094799b1595de102fa3818&csf=1&web=1&e=coUYiB) is available for reference. [Further details about processes and approvals to be added]

# What is the ISARIC SPINE?

The ISARIC4C consortium want to incorporate participants from the PHOSP, [GenOMICC](https://genomicc.org/) and [COG-UK](https://www.cogconsortium.uk/) studies into the ISARIC4C study, creating a superset of participants in a single “spine”. Doing so would allow them to match to NHS data under the ISARIC4C data agreements.

To do this, we need to have an ID and an NHS number, or CHI number, as a minimum standard. Some participants in the other studies may already be in ISARIC4C. These need to be identified and tagged as being the same individual. This will be done by comparison of NHS or CHI numbers. The spine will contain all the joins between projects and participants. Those participants that are not already in the ISARIC4C system will have an ISARIC4C ID allocated to them.

The data from the spine will be used for several purposes:

* it will feed into the NHS Digital data matching system. This system requires ID, NHS number, date of birth and postcode (the latter two are optional). This will allow the ongoing monthly matching of study participants to English & Welsh Health Data for the superset of ISARIC4C participants.
* it will allow study participants to be identified against a cohort of NHS or CHI numbers for individual research studies, and as requested by Trusted Research Environments (TREs).

Currently the data for participants in these studies are held in a series of REDCap databases. These are the source of truth for the data and update on a regular basis. We will therefore need a process to update the spine from these sources.

Source: AL, updates from JH, MC

    Kenny, clarification: Everyone in the study will be added to the spine, those without infectious diseases are still relevant as controls.

Does the API transfer need to be a separate section? How will be it relevant to this manual?

ODAP query API;  
Moni, question: Is the ODAP API a dependency for the Spine?  
Rob: No, they’re related but separate concerns.

# How do I add my dataset (within scope) into ODAP?

Check the scope

Within scope, arrange a discussion by emailing: [ODAP@ed.ac.uk](mailto:ODAP@ed.ac.uk)

Speak with the ODAP team

Agree what data and where it needs to be hosted: clinical consented data SNSH TRE and/or analysis data ODAP TRE

Does the dataset need to be linked with other datasets? Scottish health data; English health data

Can the dataset be added to the ISARIC Spine?

ODAP team will advise on most appropriate agreement: joint data controller; data processor

Agreements executed/signed off

Data to be transferred into ODAP and added to spine and/or

Refer to the the ICODA due diligence: [https://icoda-research.org/hdr-uk-as-data-controller-for-icoda, for incoming: 1](https://icoda-research.org/hdr-uk-as-data-controller-for-icoda,%20for%20incoming:%201). Datasets 2. Legal basis for use 3. Ethical review 4. Public and patient review

# How do I access ODAP for data analysis?

We absolutely must include both sides of the platform in all communications/diagrams. If we don't people get confused about what ODAP is.

So can you include another side for the embargo area? Most data will enter via it.

Access to any data in embargo section is filtered by secretariat (to remove out of scope) and then approved by the data contributor. That's the entire process. For ISARIC4C that's Kenny and Calum as joint CIs. For the other studies it is different people.

The process in the embargo area is simple:

1. Request for collaborative analysis approved by data contributor.
2. Access provided.

Under consortium arrangements (ISARIC4C; PHOSP; COG-UK; GenOMINCC)

Write it out: [ISARIC4C Data Access Process.pdf](https://uoe.sharepoint.com/:b:/r/sites/ISARIC4C/DataInfrastructureAndGovernance/IDAMAC/ISARIC4C%20Access/ISARIC4C%20Data%20Access%20Process.pdf?csf=1&web=1&e=32JQ6o)

Non consortium arrangements: ODAP process – in development (see draft below)

Data flows/Assumptions **(to be read in conjunction with flowchart):**

* In Trusted Research Environment (TRE) there are three areas:
  + Restricted area: This contains sensitive data, e.g., data under a time embargo, an active trial and/or identifiable patient data. We will not be granting researchers access to this data.
  + Embargoed data: This contains pseudonymised patient level data curated and formatted to be used for research.
  + Published data: This is the outputs from research and will contain summary/aggregate non-identifiable data only. This will be available in an open access format.
* In the Flexible Compute Space (FCS) there are two areas:
  + Embargoed data: This contains pseudonymised patient level data curated and formatted to be used for research.
  + Published data: This is the outputs from research and will contain summary/aggregate non-identifiable data only. This will be available in an open access format.
* There will be a secure API in place from the Flexible Compute Space to External TREs and UK Public Health Agencies. This is to enable data linkage across multiple data sets using unique identifiers to facilitate secure transfer if specified data fields e.g., viral sequencing.

Single Data Controller Model **– for onboarding and approval access only. – Embargo Area.**

1. The Lead Institute will become a Data Controller to facilitate the onboarding of key data sets to the ODAP and to facilitate external access to ODAP only.
2. All Data Contributors (those who hold the key data sets) will sign a Data Contributor Agreement (DCA) with the Lead Institution for the following purposes:
   * To allow the Lead Institute to onboard, curate and link the data using ODAP.
     + The Lead Institution will manage process to enable access to data however Data Contributor must give permission.
   * This Data Contributor Agreement will require a due diligence process to ensure any data can be accessed by researchers for the agreed purposes. These agreed purposes will clearly articulated and decided with public/patient input.
3. This access facilitated through the Gateway using the 5 safes data access form.
   * This governance structure sets out appropriate safeguards for privacy and effective use of sensitive data when it is embedded throughout each stage of a project
4. The Lead Institution + Data Contributor can together approve researchers access to data on ODAP if:
   * Research is for already agreed purpose. There will be specific research purposes detailed in the DCA which will all relate to patient outcomes.
   * Any linkage has been detailed in the methodology and is in an approved format. This may include only exporting certain agreed data points such as a co-morbidity score, vaccination status, covid-19 severity score, and/or outcomes.
   * All aspects of the project satisfies the 5 safes.
5. If a request is outside of the agreed purposes the Lead Institution will go back to Data Contributor and request permission.
   * This should be done either through a Data Access Committee formed between the Data Contributors or by the relevant Data Contributor only. The DAC should include public/patient representatives.
   * New project approvals will always be subject to the Data Contributors approval and a DAC should be convened of Data Contributors whose data is involves.
6. The Lead Institution will sign a Data Access Agreement with any Institutions accessing the data).
7. All aspects of any projects will be subject to clear, transparent and detailed processes.
8. The Data Contributor will remain a Data Controller of their data sets. Lead Institution becomes independent controller for facilitating access and onboarding to ODAP only.

Source: EM

## Example text for COG-UK viral sequencing data within the UKHSA Data Controller agreement

Schedule 5 Research Environment Process, Approvals and Publications

Any reference to OPAP, ODAP Internal Triage Secretariat, ODAP Data Management and Monitoring Team, OPAP Data Access Committee in this Schedule 5 shall be deemed to be a reference to Edinburgh/ Data Recipient.

1. BACKGROUND

Genomic data has been collected from SARS-CoV-2 positive samples to support the pandemic response and is freely available as an anonymised dataset. Detailed research analysis can be performed on this information if it is linked to other datasets such as vaccination status, hospital & clinic severity information, prescription data and human genome profiles. We propose to release SARS-CoV-2 genomic data from English residents (or people tested in England) to the University of Edinburgh EPCC Trusted Research Environment (TRE) where linkages to other datasets can be performed via the ISARIC4C identifier.

1. Approvals

**The Outbreak Data Analysis Platform (ODAP) NATIONAL SAFE HAVEN VS FCE:**

* The EPCC hosts the following TREs:
  + The National Safe Haven managed by PHS that holds the sensitive clinical data including from NHSD
  + FCE that will hold genomic data from UKHSA/COG
* Plan for accessing genomic data will be depend on where analysis will be performed:
  + A: Analysis using clinical / patient level data without individual patient consent:
    - Perform analysis in National Safe Haven
    - The joint data controllers (The ODAP Internal triage secretariat & UKHSA) manage access / approvals
  + B: Analysis using genomic information in the FCE either unlinked, or linked to research datasets with individual patient consent (e.g. GenOMICC: https://genomicc.org):
    - Analysis performed in the FCE
    - The joint data controllers (The ODAP Internal Triage Secretariat & UKHSA) manage access / approvals

**APPLICATION:**

1. Details of the dataset will be published on the ODAP website; HDR UK Innovation Gateway and/or the data controller’s website. HDR-UK will ensure that appropriately trained researchers (and not just COG-UK or ISARIC4C) can easily access data.
2. Researchers will submit initial data request via the HDR UK Innovation Gateway.
3. Applications will be discussed informally with (i) the data processor (ODAP Data Management and Monitoring Team), and (ii) with the data controller (i.e. UKHSA).
4. Applications will complete the full ‘Five Safes’ form and submits it via the HDR UK Innovation Gateway.

**TRIAGE/APPROVAL**:

1. The ODAP Internal Triage Secretariat (i.e. the ODAP Data Management and Monitoring Team) will review and triage applications based on the ‘Five Safes’ form. If required, data requests will be discussed with requestors.
2. Data access approval is required from two sources:
   1. The ODAP Internal Triage Secretariat or the ODAP Data Access Committee
   2. The data controller (i.e. UKHSA)
3. The ODAP Internal Triage Secretariat or the full OPAP Data Access Committee, when deemed necessary for complex requests, shall approve or decline requests on behalf of ODAP. The OPAP Data Access Committee should always have sight of data access requests even if only for information.
   * The data controller (i.e. UKHSA) will be sent details of all ODAP approved requests. Where UKHSA is joint controller, requests will be sent to a central UKHSA coordinator generic in-box; [Covid19genomics@phe.gov.uk](mailto:Covid19genomics@phe.gov.uk).
   * Where UKHSA is a joint controller, UKHSA coordinator arranges the UKHSA response on whether approval is granted for specific projects to access genomic data.
     + The response will be made within 7-days
     + The UKHSA decision will be based on joint opinion of:
       - Ian Harrison (Genomics data product owner)
       - Sam Organ (Joint Chief Data Officer)
       - Saheer Gharbia (Genomics Board)

**GOVERNANCE & LOGISTICS:**

1. The ODAP Data Management and Monitoring Team initiate process for data access with eDRIS National Safe Haven) and/or Edinburgh (FCE)
2. The ODAP Data Management and Monitoring Team
   1. Prepares data sharing agreement between the requestor and data processor/controller and relevant contracts team
   2. Ensures appropriate sign-off.

**DATA RELEASE FROM RELEVANT TRE:**

1. The ODAP Data Management and Monitoring Team will review figures and outputs of data analysis performed in the TRE before they are removed from the TRE .
2. For figures and outputs created using genomic data, the data controller (i.e. UKHSA) will also need to review and authorise removal of data.
3. DATA CREATION PROCESS

**LINKAGE IDENTIFIER USED IN TRE (NATIONAL SAFE HAVEN AND FCE)**

* ISARIC4C-ID will be used as the unique identifier. ALL COG participants will be enrolled into ISARIC4C.  This means that the ISARIC4C-ID can be used and we can link to NHSD health records
* Data compartmentalised within EPCC prevents data being linked to other datasets held by EPCC without approval. EPCC environment prevents copying or removal of data from the relevant TRE.

**Data Refresh rate:**

* Weekly refresh of the full dataset.

**GENERATION OF GENOMIC DATASET**

* CLIMB-COVID (Cloud Infrastructure for Microbial Bioinformatics) holds the genomic sequence of all the SARS-CoV-2 samples that have been sequenced in the UK, along with basic metadata, under a pseudonymised COVID-19 Genomics UK Consortium identifier (COG-ID).
* For cases registered in England or who were tested in England, UKHSA hold the information linking the COG-ID to the public health records.
* Edinburgh , in its FCE, holds multiple disparate datasets identified via an anonymised linkage key ‘ISARIC4C’ identifier.
* Public Health Scotland (PHS)/eDRIS hold the information linking the ISARIC4C-ID to an NHS number.

**Data processing:**

* UKHSA will process a defined dataset from CLIMB-COVID to replace the COG-ID identifier with ‘ISARIC4C-ID’.
* To generate the ISARIC4C-ID, UKHSA will share the NHS numbers of the cohort of individuals with PHS/eDRIS. They will generate the ISARIC4C-ID and store the linkage information mapping the identifier to the NHS number.
* Information on the linkage between NHS number and ISARIC4C-ID will be returned to UKHSA.
* UKHSA will re identify the CLIMB-COVID dataset
  + The COG-ID will be removed and replaced with a person-level identifier – the ISARIC4C-ID
  + UKHSA will add a specimen-level identifier in case individuals have had multiple samples sequenced
  + The de-identified dataset will be shared with the Edinburgh’s FCE and PHS’s National Safe Haven via secure sFTP data transfer.
* Within the National Safe Haven, the SARS-CoV-2 genomic data can be linked with other disparate clinical and epidemiological datasets, to support research studies. Linkage to other datasets is based on the ISARIC4C-ID.
* Access to the genomic dataset in the FCE will be dependent on an approval process that includes UKHSA authorisation.

1. Publication

* Publications and citations must follow the principles outlined in the COG publication policy.
* Publications should appropriately acknowledge contributions made by UKHSA & COG-UK.
* The Joint Data Controllers shall agree a Publication Policy in terms consistent with the publication policy for the ISARIC4C Study forming part of the Protocol

Project specific arrangements (include/exclude: check if data access team have started this)

To add named individuals to to project folder 1920-0273, they first need to be added to the PBPP application via an [amendment](https://www.informationgovernance.scot.nhs.uk/pbpphsc/wp-content/uploads/sites/2/2021/07/PBPP-Amendment-Request-form-v2.4-2.docx) which eDRIScan approve and notify PBPP of the additional researchers. For their access they must have completed one of the approved PBPP IG training, which is usually the MRC course. Please send copies of the MRC quiz and screenshot of completion of the modules with the amendment. To set up their safe haven accounts, please provide the below:

* Mobile numbers
* Read and return fully signed [eDRIS User Agreement](https://www.isdscotland.org/Products-and-Services/eDRIS/_docs/eDRIS-User-Agreement-v16.pdf?22)

To copy scripts from one project space to another, this is something I can do and there is no requirement to fill out any forms. Let me know if they are ready so I can copy them across so they are there for when you get access to 1920-0273.

# How will I know when data I require for my analysis is available? And where it can be found?

Need to add ideas around broad communications:

HDR UK Innovation Gateway

ODAP wiki

ODAP website/alerts

Email alerts

Ideas around data extracts communications?

ISARIC4C data file imports: via Lucy Norris (how communicate – see above)

VOC data

NIMS data

# How can I transfer my data files…? (not outputs but for further analyses)

Refer to: [ISARIC4C Data Transfer Processv1.docx](https://uoe.sharepoint.com/:w:/r/sites/ISARIC4C/DataInfrastructureAndGovernance/IDAMAC/Documentation/ISARIC4C%20Data%20Transfer%20Processv1.docx?d=wd5006b07d2da43d4a511a8528fb052a0&csf=1&web=1&e=TPoXoG)

# How can I get outputs from ODAP?

Refer to:

[NA National Safe Haven Statistical Disclosure v5.docx](https://uoe.sharepoint.com/:w:/r/sites/ISARIC4C/DataInfrastructureAndGovernance/IDAMAC/Documentation/NA%20National%20Safe%20Haven%20Statistical%20Disclosure%20v5.docx?d=w8e62494d77674ec99dd25a5b3eed5844&csf=1&web=1&e=NMtbB8)

[FCS Output Request Form - 20220608.docx](https://uoe.sharepoint.com/:w:/r/sites/ISARIC4C/DataInfrastructureAndGovernance/IDAMAC/Documentation/FCS%20Output%20Request%20Form%20-%2020220608.docx?d=w4c6e0d65713742ff92f4464c98053a99&csf=1&web=1&e=5OgOQQ)

# How is ODAP communicated? Where can I refer people to about ODAP?

See also chapter 13 (may be merge the two chapters?)

Discoverability:

HDR UK Innovation Gateway

* CO-CONNECT:
  + CO-CONNECT is HDR UK programme: limited number of cohorts from ISARIC4C