

# Cohort Discovery Implementation Guide v3.4

## Introduction

### Architecture Summary

Figure 1 - Cohort Discovery Federated Platform Architecture (50,000 ft view)

### Mapping to OMOP

Figure 2 - Example architecture separating OMOP mapping from ETL

Data Profiling and OMOP Mapping Support

Table 1 - Data profiling, OMOP mapping and ETL options

### Query Retrieving / Running Software

Table 2 - Query Retrieving Software options

### Governance Summary

Table 3 - Governance controls summary

## Onboarding Instructions

### Cohort Discovery Workstreams

Figure 3 - Cohort Discovery workstreams and tasks

Figure 4 - Interdependencies between tasks for onboarding to Cohort Discovery

### Governance Workstream

#### Governance Onboarding Steps

G1. Obtain Data Controller Consent

G2. Carry out a Data Protection Impact Assessment (DPIA)

G3. Carry out Local Approvals/Processes

G4. Technical Risk Assessment

#### Data Governance and Security Controls

Table 4 - Cohort Discovery Data Governance and Security Controls

### Data Workstream

#### Data Onboarding Steps

D1. Extract a subset of your raw identifiable data for mapping to the minimum OMOP common data model.

D2. Map the extracted subset to the OMOP common data model standard.

D3. Bonus: create synthetic data (to be used to test the ETL process in safe ways).

D4. Extract Transform Load (ETL) the extracted subset to create a new OMOP database (likely to be located on the newly created secure network area to be consumed by the query retrieving software).

### Infrastructure Workstream

#### Infrastructure Onboarding Steps

I1. Setting up a Secure Network Area

Data Partners using BC|LINK

Table 5 - Virtual Machine Minimum and Recommended Specifications

Data Partners using Bunny

I2. Install Query Retrieving / Running Software

Option I2a Install BC|LINK

Option I2b Install Bunny

Option I2c - Build it yourself

I3. Connect OMOP Data to Query Retrieving Software

## Additional Features

### Iterative Onboarding of Datasets and Fields

Onboarding multiple data cohorts

Onboarding new data fields to an existing cohort

### Enhanced Functionality

Federated Analytics

## Contracts

### Key Terms

 [Related articles](#)

## Introduction

This guide outlines the steps for Data Partners to enable their datasets to be discoverable via the Cohort Discovery tool. A “Data Partner” is any organisation onboarding data onto the tool, and is often also the controller of the data, but not always.

To understand how the Cohort Discovery tool works we highly recommend you read the introductory page on the Health Data Research

Gateway website (the Gateway) and watch the embedded video on YouTube.



### Cohort Discovery - About - Health Data Research Innovation Gateway

Cohort Discovery follows the five safe principles to ensure the safe and secure access to information about the available data across a range of research environments. The framework focusses on five key areas:



[www.healthdatagateway.org](http://www.healthdatagateway.org)

## Cohort Discovery on the Health Data Research Innovation Gateway



Each Data Partner will be introduced to key contacts within the team to guide them through the onboarding process. The [Health Data Research \(HDR\) UK](#) technology team will provide a range of supporting governance, data and infrastructure capabilities. To provide this support, the HDR technology team will not need to see your row level data. Throughout this document, “we” and “the team” are used to refer to the HDR Technology Team.

General enquiries can be emailed to [support@healthdatagateway.org](mailto:support@healthdatagateway.org)

The Cohort Discovery tool was initially integrated into the Gateway through the CO-CONNECT project. To understand more about the CO-CONNECT project, see either of the two links below:

- [A Hybrid Architecture \(CO-CONNECT\) to Facilitate Rapid Discovery and Access to Data Across the United Kingdom in Response to the COVID-19 Pandemic: Development Study - PMC \(nih.gov\)](#)
- <https://co-connect.ac.uk/>

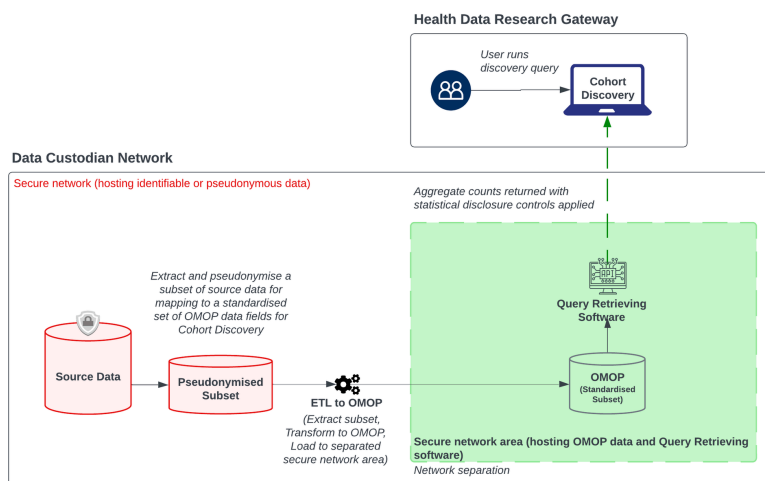
## Architecture Summary

The Cohort Discovery tool utilised by HDR UK is configured for data in the OMOP Common Data Model format. This means that a Data Partner needs to map their source data (or a subset thereof) to the OMOP common data model, and host the new OMOP data model in a secure location that can be accessed only by the Cohort Discovery query retrieving software. A secure network area is set up by the Data Partner which is separate from the location where identifiable data is stored, but still part of their secure infrastructure\* (see qualifying note below).

This secure network area should host only two things: the [minimum pseudonymised OMOP data](#) required for Cohort Discovery (see Appendix 1) and the query retrieving software that routinely pulls queries from the Cohort Discovery application.

A high-level platform architecture is below in figure 1.

Figure 1 - Cohort Discovery Federated Platform Architecture (50,000 ft view)



**i** The separated secure network area can be a Virtual Machine (VM), on-premises or cloud infrastructure, and needs to support Kubernetes container orchestration.

\*NB: as an alternative, the system can technically be securely figured so that the Query Retrieving Software does not sit within a separate VM, and/or the OMOP pseudonymised data connected is simply a view on data stored within a larger datastore. However, our experience is that clear separation of the zones and data provide additional comfort to Data Governance panels regarding the technical controls and therefore we recommend consideration of this when deciding on the architecture.

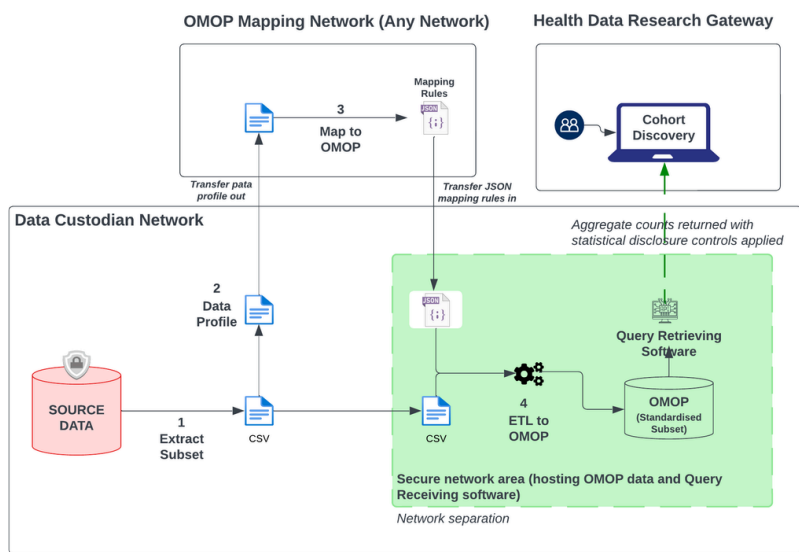
## Mapping to OMOP

**i** See Appendix 1 for the [minimum set of OMOP data fields required for Cohort Discovery](#)

Mapping source data to OMOP can be challenging and requires specialist skills and tools. This can all be done in one ETL process, however it can be easier to separate mapping logic from ETL. To separate mapping logic from ETL, the example architecture in figure 2 below outlines a way to do this without compromising sensitive data.

In figure 2, each Data Partner extracts a subset of their data [1], profiles the data [2], maps the profiled data to OMOP (creating mapping rules) [3], and then transforms their extracted data (from step 1) to OMOP using the mapping rules (from step 3), and loads it onto the virtual machine [4] where it is made available to query by the query retrieving software.

Figure 2 - Example architecture separating OMOP mapping from ETL



Data Partners can develop their own processes for data profiling [2], OMOP mapping [3] and ETL [4], and use a combination of commercial and open source software.

Data Profiling and OMOP Mapping Support

The OMOP and open-source community have developed tools to support the steps of data profiling, mapping to OMOP and ETL to OMOP at a target destination. HDR UK and partners can offer support with the tools below:

Table 1 - Data profiling, OMOP mapping and ETL options

Process	Tool	Description
Data Profiling	White Rabbit	Developed by <a href="#">OHDSI</a> to help prepare ETL of longitudinal healthcare databases into the OMOP Common Data Model (CDM). The main function of White Rabbit is to perform a scan of the source data, providing detailed information on the tables, fields, and values that appear in a database. This tool is used for structural mapping. White Rabbit is typically the first piece of software used in the end-to-end ETL process.
Map to OMOP	Carrot Mapper	Carrot Mapper is a web-tool originally designed and developed by the <a href="#">CO-CONNECT</a> project team that enables the Data Partner to map the White Rabbit output and generate a "Mapping File" in JSON format. This mapping file defines the guidelines for the ETL process on the dataset(s). See the <a href="#">Carrot Documentation</a> for instructions on how to install and run Carrot Mapper: 🐰 <a href="#">Carrot-Mapper - Carrot Docs</a>
ETL to OMOP	Carrot-CDM	Carrot-CDM is An ETL tool designed and developed by the <a href="#">CO-CONNECT</a> project team. This tool automates the extraction of pseudonymised data, the transformation of data to the OMOP CDM and the loading of this data to the Query Retrieving Software. See the <a href="#">Carrot Documentation</a> for instructions on how to install and run CaRROT-CDM: 🐰 <a href="#">Carrot-CDM - Carrot Docs</a>

Note that other tools are available for all of those processes, for example [Usagi](#) is developed by [OHDSI](#) to aid the manual process of creating OMOP code mappings, however HDR UK has not worked with these tools previously and therefore cannot offer any support.

📢 HDR UK and partners at the Dundee Alleviate Hub offer OMOP mapping services using CaRROT tools (other commercial vendors also exist). Please reach out to your HDR UK contact for more information.

Query Retrieving / Running Software

This software routinely pulls queries from the Cohort Discovery application to run against the OMOP data. A Data Partner can build their own tool to perform this capability using [this API documentation](#) or they can use one of the commercial or open source options in table 2 below.

Table 2 - Query Retrieving Software options

Tool	Type	Developed By	Description
------	------	--------------	-------------

BC LINK	Commercial	BC Platforms	<p>BC LINK is free to use and can be installed on the Data Partner's secure infrastructure. It runs on PostgreSQL (a relational database management system) which hosts the Data Partner's pseudonymised OMOP data.</p> <p>End users submit queries via the Cohort Discovery portal on the Health Data Research Gateway, BC LINK retrieves query requests from the portal and translates them to SQL which is then run against the pseudonymised OMOP database.</p> <p>Summary statistics are returned to the end user via the query portal. These are controlled for statistical disclosure by low number suppression and rounding (set by the Data Partners).</p> <p>Data Partner's using BC LINK will be asked to sign a license agreement with BC Platforms for the free use of their software.</p> <p>HDR UK can provide some basic support for this tool, escalating to BC Platforms if there are any bugs found. BC Platforms may be able to offer an additional support contract on request if required.</p>
Bunny	Open Source	University of Nottingham	<p>Bunny is an open-source application that supports Cohort Discovery. Bunny fetches Cohort Discovery queries and resolves them against an OMOP database.</p> <p>Bunny is deployed in your local environment and makes only outgoing requests, which safely enables queries to be executed behind your firewall. Bunny enables obfuscation of query results to simplify data governance issues, and can be part of a federated network through Hutch Relay.</p> <p>The Bunny user guide is available here: <a href="https://health-informatics-uon.github.io/hutch/bunny">https://health-informatics-uon.github.io/hutch/bunny</a></p>

**i** HDR UK and partners can support Data Partners who wish to use Bunny. Please reach out to your HDR UK contact for more information.

## Governance Summary

The architecture has been designed to allow Data Partners to retain full control of their data and to simplify data governance requirements. Both BC|LINK and Bunny provide the same functionality to apply the controls. The following controls are in place to protect patient confidentiality and to ensure the security of the data:

Table 3 - Governance controls summary

#	Data Protection Mechanism	Description
1	Raw data access	Identifiable data is only handled by the Data Partner and only within their own secure environment. The Data Partner will pseudonymize their data before it is transferred to a secure network area within their own environment, and they have full control of which aspects of their data are uploaded to Cohort Discovery.
2	PII removal	All Personally Identifiable Information (PII) (i.e., patient location, date of birth, names etc.) relating to subjects in the datasets is to be withheld; no PII will be held in any software.
3	Disclosure control	<p>Only aggregated results for any discovery query will leave a Data Partner's control, subject to automated disclosure controls. Data Partners can control the low count suppression, for example counts of less than 10 are returned as 0 and counts above 10 can be rounded off to the nearest 10.</p> <p>In addition to the rounding of results, it is not possible to query the ID of an individual person in the Cohort Discovery query builder. This means that "differencing attacks" are not possible.</p>
4	Query control	<p>Queries run by end users (i.e., researchers) of the Cohort Discovery application (BC Platforms RQUEST software embedded within the Health Data Research Gateway) can only be constructed from pre-defined fields and they can only query data that has been authorized via a drag &amp; drop interface. Data Partners set both the threshold at which no results are returned and whether rounding should be applied. This is a configuration of either the BC Link or Bunny software.</p> <p>Users are also limited in the number of queries they can run repeatedly which minimizes the risk of re-identifying individuals.</p>

HDR UK will also utilise the BC|REQUEST API to extract summary level demographic information to display in the Gateway, alongside the already published metadata for a given data collection, to enhance information available to researchers. Such demographic information will include counts by age distribution, ethnicity, sex and diseases. This will be controlled for statistical disclosure.

**i** Data Partners are advised to carry out a Data Protection Impact Assessment (DPIA), although it is expected that as the data is anonymised with appropriate controls the DPIA will return that GDPR does not apply (this is our experience to date). The HDR UK technology team can support the drafting of DPIAs.

## Onboarding Instructions

Most Data Partners are likely to have already contributed metadata on the Gateway.

**i** The Gateway includes a metadata catalogue from across multiple datasets and enables them to be searchable and accessible. Depositing metadata in the Gateway will enable researchers and users to understand the data available, although it does not contain any source data such as record counts. The Gateway will help the researchers and innovators by providing:

- Quick access to the metadata by searching through keywords, i.e., phenotypes, coverage etc.
- A location for the data
- Details of the dataset held including field names and descriptions as well as data schemas.

The Cohort Discovery tool provides a link to the Gateway metadata to enable users to find out additional details about a dataset, e.g., contact details. Therefore, it is highly recommended that a metadata record is also available on the Gateway to associate with the Cohort Discovery tool. This will facilitate a smoother journey from data discovery through to data access requests.

For Data Partners that have not contributed metadata to the Gateway, here's how: <https://discourse.healthdatagateway.org/c/how-to/data-custodians/23>.

**i** Please note that HDR UK are currently working on an enhanced version of the Gateway with expected launch in September 2024. Therefore, please reach out to the technology team to discuss whether it might be more appropriate to upload data utilising the enhanced version rather than via the current one.

Cohort Discovery and Gateway implementations can occur simultaneously. The rest of this guide will describe the steps to follow for Cohort Discovery implementation only.

## Cohort Discovery Workstreams

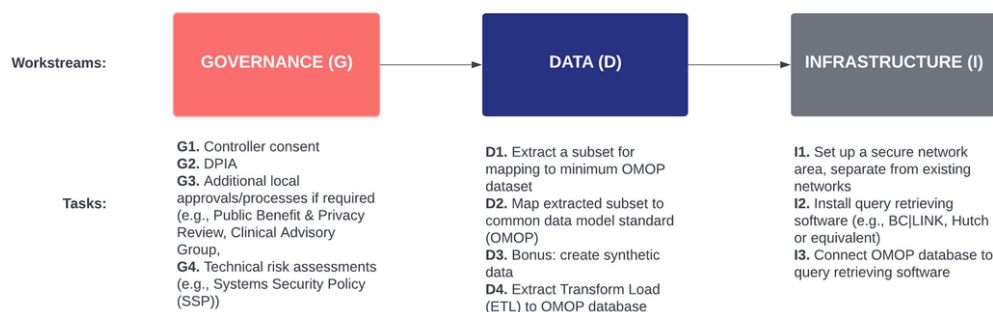
At the start of the onboarding process, an initial joint meeting with the HDR UK technology team and the Data Partner will be held to answer any questions and to define the exact tasks and responsibilities from that point forward.

There are three core workstreams to complete to onboard your data to Cohort Discovery:

1. Information Governance (IG)
2. Data
3. Infrastructure

Each workstream contains a number of core tasks (see figure 3, with interdependencies between tasks in figure 4 below):

Figure 3 - Cohort Discovery workstreams and tasks



**i** IG processes should be started as soon as the project is initiated. Workstreams can run in parallel, although there are some interdependencies between some tasks (see figure 4 below).

**Figure 4 - Interdependencies between tasks for onboarding to Cohort Discovery**



Data Partners are likely to also be the Data Controllers. If a Data Partner is not a Data Controller, consent must be obtained from the relevant Data Controller(s) regarding the use of the data. The Cohort Discovery architecture is designed on a flexible technical philosophy that means Data Controllers can participate with their current governance and consent.

Note that governance steps are not always linear and can occur in any order.

**Responsible for this action:** Data Partner (HDR UK technology team can support)

### Description

This needs to be completed before you can connect the 'live' data and query receiving software to either the 'test' or the 'live' BC|REQUEST software. You can still complete the data steps D1 (extract a subset of your data for mapping to the minimum OMOP variables), D2 (mapping your extracted subset to OMOP), and D3 (creating a synthetic dataset for testing the ETL process), but you cannot load any real data onto the secure network area to the query receiving software if it is connected to the Cohort Discovery tool.

 Only synthetic data should be used to test the system until G1 is complete.

### Description

A DPIA is a process to help you identify and minimise the data protection risks of a project. You must do a DPIA for processing that is likely to result in a high risk to individuals. See the [Information Commissioners Office \(ICO\) for more details on data protection impact assessments](#).

**i** While it is expected that the DPIA assessment will conclude that the participation in Cohort Discovery is minimal risk and that the data to be onboarded is sufficiently anonymized to the point where it falls outside the remit of GDPR, it is important that Data Partners make that conclusion independently.

### G3. Carry out Local Approvals/Processes


### Description

Data Partners may decide that additional protocols are necessary, such as

1. Public Benefit & Privacy Review (e.g., the Public Benefit & Privacy Panel in Scotland)
2. Internal data access applications

#### G4. Technical Risk Assessment

Data Partners may decide that additional technical risk assessments are undertaken, such as creating and maintaining a System Security Policy (SSP).

 Typically, governance steps require knowing what data you are going to select to make available for discovery, so they often go hand in hand with step D1 - Data Preparation.

#### Data Governance and Security Controls

Cohort Discovery is designed so that data discovery is performed without requiring the data to move. Several measures are in place to protect and safeguard the data.

In addition to the steps taken in the governance workstream (G1 to G4), Cohort Discovery includes several key controls to protect patient confidentiality and data security (see table 4 below). It is recommended that information on these controls are included within DPIAs.

Table 4 - Cohort Discovery Data Governance and Security Controls

#	Control	#	Sub-Control	Description
1	Raw Data Access	a	Handling	<b>Identifiable data is only ever handled by Data Partner</b> employees within the Data Partner IT infrastructure.
		b	Pseudonymisation	<b>Data is pseudonymised and all personal identifiable information relating to subjects is removed by the Data Partner</b> before being connected to the Cohort Discovery query receiving software.  Pseudonymised record level data will not be available outside of the Data Partners infrastructure.
		c	PII removal	<b>All Personally Identifiable Information (PII) relating to subjects in the datasets is to be withheld</b> (i.e., patient location, date of birth, names etc.); no PII will be held in any software outside of the raw data store.
2	Firewalls			<b>Firewalls</b> will be in place with standard controls against malicious attacks and require no additional inbound rules.
3	Pseudonymised Data Access (OMOP)	a	Cohort Discovery (BC RQUEST) Accessibility	The Cohort Discovery tool (which uses an application called BC RQUEST) is accessible to end users <b>only via the Gateway</b> . End users, HDR UK employees or BC Platform employees will not be able to directly access the row level pseudonymised OMOP data hosted by the Data Partner.
		b	Disclosure Control	<b>Only metadata and aggregated</b> results for any query will leave the Data Partners control. The Data Partner can control what goes out through two key controls (both configurable by the Data Partner):  i. <b>Low count suppression</b> - for example, counts of less than 10 are returned as 0  ii. <b>Rounding</b> - counts between 5 and 10 are rounded off to the nearest 10  In addition to the rounding of results, it is not possible to query the ID of an individual person in the Cohort Discovery query builder. This means that “differencing attacks” are not possible.
		c	Query Control	End user <b>queries can only be constructed from pre-defined fields</b> within BC RQUEST. The user interface ensures that users can only query data that has been authorised.  Note that the HDR UK technology team will query the Cohort Discovery tool via an Application Programming Interface (API) in order to extract high level demographic frequency distribution information such as age, sex, ethnicity and diseases, to allow for updating existing Data Partner information on the Gateway. This will enable researchers to understand more properties about the Data Partners datasets from within the Gateway, without out having to log on to the Cohort Discovery tool and will also help the Gateway to return more accurate searches.



		<p>The API will only be queried by the HDR UK technology team and the query receiving software (i.e., BC LINK or Bunny or any additional application implemented by the Data Partner) and will not be open for any other querying (i.e., end users will not be granted access to call these API's directly in their analysis scripts).</p> <p>HDR UK will also query BC RQUEST application logs to provide additional transparency to Data Partners on who is using the application to access their data, beyond what is routinely provided within the application itself.</p>
d	Query Logging	<b>All queries are logged</b> and can be reviewed by the Data Partner from the log management software.
e	Authentication I	<p>All users of Cohort Discovery <b>must firstly be authenticated users of the Gateway</b>. Anyone can register with the Gateway (via Google, Microsoft, LinkedIn, Orcid or via OpenAthens), however Gateway users who also wish to use Cohort Discovery undergo additional authentication.</p> <p>Requests for access are assessed in accordance with the <b>UK Health Data Research Alliance</b> principles for participation using a proportionate governance approach based on the <b>Five Safes Framework</b>. For Cohort Discovery, we focus on <b>Safe People</b> and <b>Safe Projects</b>, with the other three Safes (Setting, Data and Outputs) being managed by the data and technology partners.</p>
	Authentication II	<p>Cohort Discovery can only be accessed by <b>'Safe People'</b>. Currently, every application to access the Cohort Discovery tool is validated manually by the HDR UK administration team. The criteria for a Safe Person can be determined by each Data Partner. Users of Cohort Discovery are only able to run queries against those datasets where they have passed the Safe Person criteria of that dataset. This is achieved by a configuration of BC RQUEST which is carried out by the HDR UK administration team when a new user is set up.</p> <p>As an example, one Data Partner might only permit queries on their data via the Cohort Discovery tool from researchers who work in a UK academic institution, and are located within the UK. A new user, who is not from a UK academic institution or is located outside of the UK will be configured so that they cannot query data from this Data Partner.</p> <p>These checks will ideally be standardised across Data Partners where possible, and HDR are continually working with Data Partners to agree these authentications.</p> <p>Generally, Safe People must:</p> <ul style="list-style-type: none"> <li>• Work for appropriate and trusted organisations from either academia, public sector or industry. Cohort Discovery does not accept applications from generic email domains (such as gmail, hotmail etc.).</li> <li>• Demonstrate their role as a bona fide researcher, NHS analyst or equivalent. A person may receive bona fide researcher status if their host organisation confirms they are a current researcher and the host organisation is a valid organisation undertaking research.</li> </ul> <p>The checks currently undertaken by HDR UK include:</p> <ul style="list-style-type: none"> <li>• Checking a user's organisational email address is active by an email exchange.</li> <li>• Checking a user's publication record, e.g., checking their ORCID.</li> <li>• General internet checks on an individual's profile.</li> <li>• Checking the organisation a researcher is from is public sector / academic (for data partners who wish to exclude industry access).</li> </ul> <p>HDR UK are in the process of developing a researcher registry with a working title of 'Safe Organisation and User Registry for Sensitive Data (SOURSD)'. A UK-wide standard is being developed through the Pan-UK Data Governance Steering Committee (part of the Health Data Alliance). HDR UK are also developing a technical implementation of the standard. This solution will enable a representative from the SDE network to assess a researcher who requests access to the Cohort Discovery tool and record within the registry if access is approved or denied. The Cohort Discovery tool is being enhanced to query the registry to assess if access is approved to query the SDE Network data on a per user basis. A minimum viable product version of SOURSD to capture approvals is planned for April 2025.</p> <p>Prior to this automated solution being in place, HDR UK could carry out researcher checks (as is the case for other Data Partners) or a representative</p>

from the SDE Network could carry out the checks and then inform HDR UK of the decision (HDR UK would then configure access accordingly).

**However, it is HDR UK’s preference that the SDE Network are responsible for the review of users and approving access.**

HDR UK recommend that pre-April 2025, a process is set up whereby:

- HDR UK emails an SDE representative with the details of a new user request.
- The SDE network representative carries out the relevant checks (to be agreed by the SDE network) and then emails HDR UK to inform them of the decision.
- HDR UK then informs the potential user of the outcome and configures the system access accordingly.

Post-April 2025 (when Researcher Registry is available):

- The Cohort Discovery tool automatically informs an SDE representative of a new user request.
- The SDE representative reviews the request using information provided in SOURSD (and additional information if necessary) and then approves / denies the request for access within SOURSD.
- The SOURSD application automatically informs the researcher if access has been approved / denied.
- The Cohort Discovery tool automatically queries SOURSD to assess if access is allowed.

**Please note that this is an ongoing area of development with the SDE Network and is therefore subject to change.**

Authentication III Cohort Discovery can only be accessed by safe people undertaking ‘**Safe Projects**’ that have the potential for **public benefit**.

Currently, users of the Cohort Discovery tool have to draft a public benefit statement related to the specific queries they wish to run. It is only possible to have one public benefit statement per user and so does not enable users to use the tool to find data for multiple projects. The new major version of the Gateway to be launched in summer 2024 reduces the bottleneck for users having to draft a public benefit statement and enables them to use the tool for multiple, concurrent feasibility projects. Users are instead required to agree to the terms of use:

*“I confirm that my use of the tool will be for the sole purpose of understanding the size of populations relevant to a particular research question and that any resulting study or clinical trial will be for public benefit as defined by the [National Data Guardian](#).*

*I acknowledge that:*

1. *Any subsequent access to the data for research purposes (including publications) will be subject to the relevant data access processes of the data custodian(s).*
2. *Some data custodians do not allow results from queries on their data to be visible to commercial organisations and the tool will not show results from queries run by commercial users in these cases.*

*I will not:*

1. *Share, publish or communicate the results or findings from the tool in any way, with the exception of sharing results with collaborators working on a research study or clinical trial and in support of an application for research funding for an eligible research study or clinical trial.*
2. *Use the results for marketing or policy development.*

*I understand that my access to the tool may be terminated if I do not comply with the above.”*

Users will be required to re-sign these terms every 6 months.

**Please note that this is an ongoing area of development with the SDE Network and is therefore subject to change.**

Additionally, all users of the Gateway are required to adhere to the Gateway terms of use: [HDRUK Innovation Gateway | Terms and Conditions](#)

				( <a href="http://healthdatagateway.org">healthdatagateway.org</a> ).
4	Synthetic Data Access	a	General	<p>Data Partners who use synthetic data for testing the ETL processes and general infrastructure can make this synthetic data available for querying by users who do not fall easily into any of the desired user categories specified above.</p> <p>For example, to support ongoing development of Cohort Discovery within the wider community, we may grant access to the Cohort Discovery tool to users who are not interested in the results of queries, but only interested in observing how the tool works, for example governance staff and developers of similar products such as NHS DigiTrials. These users can be granted access to synthetic data collections only.</p>

**i** Note that some Data Partners may wish to promote their presence on Cohort Discovery through website publications, announcements and blogs. Occasionally this might result in media attention being drawn upon Cohort Discovery tool.

## Data Workstream

### Data Onboarding Steps

The data workstream contains steps that might vary depending on your chosen methods of extracting a subset of your raw identifiable data, mapping this data to the OMOP common data model, transforming the data to OMOP and loading it into your secure network area where it will be available for querying.

The core high level steps required for all methods are:

**D1. Extract a subset of your raw identifiable data for mapping to the minimum OMOP common data model.**

**D2. Map the extracted subset to the OMOP common data model standard.**

**D3. Bonus: create synthetic data (to be used to test the ETL process in safe ways).**

**D4. Extract Transform Load (ETL) the extracted subset to create a new OMOP database (likely to be located on the newly created secure network area to be consumed by the query retrieving software).**

There are multiple ways a Data Partner might implement these core steps. Mapping source data to OMOP can be challenging and requires specialist skills and tools. This can be done in one ETL process, however it can be easier to separate the mapping logic from the ETL step. Separating the mapping logic from the ETL step allows non-technical subject matter experts to create mapping rules used in any ETL process.

The CO-CONNECT programme mapped source data to OMOP outside of the secure data partner network in secure ways that protected data privacy using White Rabbit for data profiling, CaRROT Mapper for mapping profiled data to OMOP, and the CaRROT CDM tool to extract, transform and load source data into a new de-identified OMOP common data model on a newly created secure network area. CO-CONNECT also used the BC|LINK application to manage queries between OMOP data and the Cohort Discovery tool.

The process and architecture for Data Partners wishing to follow this method is available in Appendix 2 - [Data onboarding steps using White Rabbit and Carrot](#).

## Infrastructure Workstream

The infrastructure workstream contains the following key implementation steps:

### Infrastructure Onboarding Steps

#### I1. Setting up a Secure Network Area

**i Responsible for this action:** Data Partner (HDR technology team available to answer any related questions but will not have direct access to the system)

**Who can see the data/process:** Data Partner

**Where does the process take place:** Within the Data Partner's environment (secure network)

#### Description

Data Partners need to provide a secure environment within their network to house and process their pseudonymised data. The ETL tool will load data into this area and the Query Retrieving Software will be installed in this area to allow queries to be performed. This way, your pseudonymised OMOP data is separated from your primary data source.

**i** Note that as an alternative, the system can technically be securely configured so that the query receiving software does not sit within a separate VM and/or the OMOP pseudonymised data connected is simply a view on the data stored within a larger datastore. However, our experience is that clear

separation of the zones and data provided additional comfort to Data Governance panels regarding the technical controls, and therefore recommend consideration of this when deciding on an architecture.

The new secure network area requires outbound connection only over HTTPS to a specific IP range in order to connect to the BC|REQUEST application.

There are several options for creating a secure environment.

#### Data Partners using BC|LINK

For Data Partners using BC|LINK as their Query Retrieving Software, the most platform agnostic and portable approach is to deploy it in a K3s Kubernetes distribution on a virtual machine. Other methods are possible, for example using cloud native providers such as Azure, Amazon and Google, however HDR UK does not currently offer support for these environments and the Data Partner will need to contact BC Platforms directly for support (BC Platforms are the vendors of BC|LINK and BC|REQUEST software).

The minimum specifications for supported virtual machines which rely on RKE2/K3s clusters running within them are provided below in table 5 (note that these are minimum specifications, BC|LINK will also work on other OS versions, for example Rocky 9).

Table 5 - Virtual Machine Minimum and Recommended Specifications

Virtual Machine Requirements	Minimum Specifications	Recommended Specifications
<b>OS version</b>	Ubuntu 22.04 LTS or Higher	N/A
<b>CPU &amp; Memory</b>	8 vCPUs, 32GB Memory	16 vCPUs, 64GB Memory
<b>Storage</b>	<ul style="list-style-type: none"><li>100GB storage mounted at root partition</li><li>400GB storage mounted at /data</li></ul>	<ul style="list-style-type: none"><li>100GB storage mounted at root partition</li><li>1TB storage mounted at /data</li></ul>
<b>Database</b>	PostgreSQL deployment within same VM is possible. <ul style="list-style-type: none"><li>Latest PostgreSQL version 15.x/16.x</li><li>LINK: Dedicated core 4vCores, 16GiB memory Storage &gt;= 500GB</li></ul>	<ul style="list-style-type: none"><li>Dedicated PostgreSQL server is highly recommended.</li><li>LINK for large clinical database &amp; small scale genome data: Dedicated core 8vCores, 32GiB memory, Storage &gt;= 2TB</li><li>LINK for large clinical database, with population-scale genome data: Dedicated core 16vCores, 64GiB memory, Storage &gt;= 2TB + 1TB for each 20K subjects with whole genome data</li></ul>
<b>Third Party Tools Required</b>	<ul style="list-style-type: none"><li>K3S</li><li>Helm</li><li>Docker (with docker compose)</li><li>Keycloak w/ Dedicated core 2vCores, 8GiB memory, 100GB Storage</li></ul>	
<b>Networking Requirements</b>	Customer/Site Specific - to be discussed	

**i** Data Partners using CentOS 7 will need to migrate to a new OS before Summer '24 as this is approaching its end of life, or purchase an end of life add on for Red Hat 6 customers. For non Red Hat users, you may also consider converting to Red Hat to give you more time before migrating.

#### Data Partners using Bunny

For Data Partners using Bunny, please see the Bunny user guide which is available here: <https://health-informatics-uon.github.io/hutch/bunny> or contact the University of Nottingham for optimal environments.

## 12. Install Query Retrieving / Running Software

**i** **Responsible for this action:** Data Partner (HDR technology team available to answer any related questions but will not have direct access to the system)

**Who can see the data/process:** Data Partner

**Where does the process take place:** Within the Data Partner's environment on a secure network area

### Description

This software routinely pulls queries from the Cohort Discovery application to run against the OMOP data. A Data Partner can build their own tool to perform this capability using this [API documentation](#) or they can use the supported commercial or open source options in table 2 above.

#### Option I2a Install BC|LINK

BC|LINK is a software application developed by BC Platforms and runs on a Postgres database. BC|LINK can be installed on the secure network area created above (in step I1). BC|LINK will connect to the database of pseudonymised data in the OMOP common data model.

Each Data Partner will be asked to configure and install the BC|LINK software. Data Partners installing BC|LINK on a virtual machine using K3s can get support from HDR technology team if required. See Appendix 3 for BC|LINK installation instructions ( BC\_LINK 6.3.x Deployment and Technical VM).

Once the BC|LINK application is installed, it will **only** communicate with the BC|RQUEST application (i.e., with Cohort Discovery) on the Gateway. This is done on the current BC|RQUEST via a secure SSH tunnel to pull down queries then run and return query results. This is an outbound connection only, meaning you do not need to allow SSH access to your infrastructure. All Data Partners will be expected to do their own security and vulnerability testing of the application and the communication channel, if required. In future versions of BC|RQUEST, this communication will be via a REST API outbound on 443 to pull down queries, then run and return query results. This is an outbound connection only, meaning you do not need to allow inbound HTTPS access into your infrastructure.

**i** The Secure Shell (SSH) protocol is a method for secure remote log in from one computer to another. It provides several alternative options for strong authentication, and it protects communications security and integrity with strong encryption. It's a secure alternative to the non-protected login protocols (such as telnet) and insecure file transfer methods (such as FTP).

#### Option I2b Install Bunny

Bunny is an open-source application that supports Cohort Discovery. Bunny fetches Cohort Discovery queries and resolves them against an OMOP database.

Bunny is deployed in your local environment and makes only outgoing requests, which safely enables queries to be executed behind your firewall. Bunny enables obfuscation of query results to simplify data governance issues, and can be part of a federated network through Hutch Relay. Container images are available for ease of deployment in your environment, and Bunny can also be ran just as a query executor through its command line interface.

The Bunny user guide is available here: <https://health-informatics-uon.github.io/hutch/bunny>.

#### Option I2c - Build it yourself

Data Partners are welcome to develop their own tool to retrieve queries from the Cohort Discovery tool and run against their OMOP datasets. Any new tools will need to meet the required API standards set out in the [Swagger documentation](#), please contact the HDR technology team to discuss this option.

### I3. Connect OMOP Data to Query Retrieving Software

**i** **Responsible for this action:** Data Partner (HDR technology team available to answer any related questions but will not have direct access to the system)

**Who can see the data/process:** Data Partner

**Where does the process take place:** Within the Data Partner's environment on a secure network area

#### Description

Once the OMOP data has been created and stored within a database, it needs to be connected to the Query Retrieving Software so that queries can be run.

1. Upload the database containing OMOP format pseudonymised data to the same network space as the Query Retrieving Software.
2. If it is not already connected, then connect the database to the Query Retrieving Software (if using BC|LINK, this is done via the graphical user interface in a web browser).
3. Test that queries can run from the Cohort Discovery tool.

## Additional Features

In addition to the three onboarding workstreams, Cohort Discovery supports the following features.

### Iterative Onboarding of Datasets and Fields


Cohort Discovery supports an iterative approach for onboarding a range of data cohorts and data fields.

#### Onboarding multiple data cohorts

A Data Partner with multiple data cohorts can deploy a single institutional BC|LINK (or Bunny, or custom) instance/server since a single BC|LINK (or Bunny, or custom) server can store and make multiple cohorts discoverable. Additional configuration steps are required to enable this capability. Data Partners with multiple data cohorts are asked to contact the HDR technology team for further support.

#### Onboarding new data fields to an existing cohort

Within each of the datasets, the most important fields will be onboarded first. If additional new data fields are identified, the process of including these additional data fields can be iterative. We recommend first focusing on the key fields required to answer pressing research questions. This process will allow the addition of new fields requested by the community to be added over time.

 The iterative onboarding process for both new collections and/or new data fields can follow the same governance process explained in Data Governance and Security Controls.

## Enhanced Functionality

### Federated Analytics

The Cohort Discovery architecture has the potential to be enhanced to support Federated Analytics. To run Federated Analytics queries and to use the resource for a full research project (rather than just a Cohort Discovery / feasibility check), a researcher would require additional data governance permissions and there may be additional associated costs required by the Data Partner.

HDR are researching how to enhance the architecture via their Federated Analytics workstream. This workstream is researching methods to support Data Partners to perform semi-automated extraction of OMOP data into Trusted Research Environments (TREs), how to provide counts of population overlaps across datasets, and a standardised method for sending queries to TREs and performing automated disclosure control. These capabilities are not currently implemented in production.

The Alleviate HDR Hub for Pain is also using the infrastructure to support federated PheWAS and GWAS queries over consented genomic data. This is a feature of the Cohort Discovery tool.

Please reach out to the HDR UK technology team if you are interested in learning more about Federated Analytics and/or utilising the tool to support federated genomic data analysis.

---

## Contracts

Data Partners opting to use BC|LINK will be asked to sign a license agreement with BC Platforms (at no cost).

---

## Key Terms

Term	Stands For	Description
BC LINK		BC LINK software is developed by BC Platforms. BC LINK Application (App) is installed on a virtual machine (VM) within the Data Partner's secure infrastructure and runs on PostgreSQL (a relational database management system) which hosts the Data Partner's pseudonymised data. End users submit queries via the Cohort Discovery portal on the HDR Gateway, the App receives query requests from the portal and translates it to SQL which is then run against the pseudonymised OMOP database. Summary statistics are returned to the end user via the query portal. N.B.: No record level information is returned, only the aggregates/summarised statistics are returned.
BC RQUEST		The centralised query portal allows end-user to enter search queries. BC RQUEST processes these queries before passing them to each of the BC LINK instances. BC RQUEST also receives back summary statistics that are returned to the end-user. BC RQUEST is developed by BC Platforms.
BC Platforms		BC Platforms is a global leader, operating out of the UK, providing a powerful, modular, data and technology platform for federated healthcare data, personalized medicine, and drug development, accelerating the translation of insights into clinical practice. They have a proven record of accomplishment in delivering automated data harmonization solutions ( <a href="#">Case Studies</a> )
Bunny		Bunny is an open-source application that supports Cohort Discovery. Bunny fetches Cohort Discovery queries and resolves them against an OMOP database.  Bunny is deployed in your local environment and makes only outgoing requests, which safely enables queries to be executed behind your firewall. Bunny enables obfuscation of query results to simplify data governance issues, and can be part of a federated network through Hutch Relay.

		<p>The Bunny user guide is available here: <a href="https://health-informatics-uon.github.io/hutch/bunny">https://health-informatics-uon.github.io/hutch/bunny</a>.</p> <p>Note that Bunny has evolved from the wider Hutch stack to specifically serve upstream Task API's such as Cohort Discovery.</p>
<b>CO-CONNECT</b>	Curated and Open aNalysis aNd rEsearCh plaTform	CO-CONNECT a research project which set up the initial Cohort Discovery infrastructure and service (ended Oct 2022).
<b>Carrot-Mapper</b>	CaRROT-Mapper	A web-tool originally designed and developed by the CO-CONNECT project team that enables the Data to map the Scan report and generate a "Mapping File" in JSON format. This mapping file defines the guidelines for the ETL process on the dataset(s).
<b>Carrot-CDM Tool</b>		An ETL tool designed and developed by the CO-CONNECT project team. This tool automates the Extraction of Pseudonymised data, Transformation of data to OMOP CDM and Loading to BC LINK Process.
<b>Data Controller</b>		<p>The natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.</p> <p>Data Controllers are responsible for complying with the GDPR and therefore must be able to demonstrate compliance with the data protection principles, and take appropriate technical and organisational measures to ensure your processing is carried out in line with the GDPR.</p>
<b>Data Curation</b>		The organization and integration of data collected from various sources.
<b>Data Custodian</b>		<p>Responsibilities for data management are increasingly divided between the business process owners and information technology (IT) departments. Two functional titles commonly used for these roles are Data Steward and Data Custodian.</p> <p>Data Custodians are responsible for the safe custody, transport, storage of the data and implementation of business rules. Simply put, Data Custodians are responsible for the technical environment and database structure.</p>
<b>Data Dictionary</b>		Information about data such as table and field descriptions, relationships to other data, origin, usage, and format.
<b>Data Discovery</b>		The process of obtaining actionable information by finding patterns in data from multiple sources.
<b>Data Governance</b>		Managing data assets throughout their lifecycle to ensure they meet organisational quality, integrity and confidentiality standards.
<b>Data Interoperability</b>		Addresses the ability of systems and services that create, exchange, and consume data to have clear, shared expectations for the contents, context and meaning of that data.
<b>Data Partner</b>		An organisation onboarding federated data into the Cohort Discovery tool
<b>Data Processor</b>		A natural or legal person, public authority, agency, or other body which processes personal data on behalf of a Data Controller.
<b>Dataset</b>		A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.
<b>DPIA</b>	Data Protection Impact Assessment	A process to help identify and assess project data risks. Data Partners must complete a DPIA for processing that is likely to result in a high risk to individuals. This includes some specified types of processing. Screening checklists are available to help decide when a DPIA is necessary. The DPIA must describe the nature, scope, context, and purposes of the processing; assess necessity, proportionality, and compliance measures;

		<p>identify and assess risks to individuals; and</p> <p>identify any additional measures to mitigate those risks.</p>
<b>ETL</b>	Extract Transform Load	A type of data integration that refers to the three steps (Extract, Transform, Load) used to combine data from multiple sources into a destination system which represents the data in a different way than the source, preparation for the ETL is supported by White Rabbit and Rabbit in a Hat.
<b>GDPR</b>	<a href="#">General Data Protection Regulation</a>	A legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU).
<b>Health Data Research UK (HDR UK)</b>		The UK national institute of Health Data Research whose mission is to unite the UK's health and care data to enable discoveries that improve people's lives by providing scalable and robust data infrastructure and services.
<b>The Gateway</b>	The Health Data Research Gateway	The portal where researchers can search, discover, and request access to datasets, tools, and resources for research purposes.
<b>Metadata</b>		<p>Metadata is information about the data and this document refers to two types.</p> <p>Structural Metadata, which gives information about the table names and field names in each table for each data set.</p> <p>Descriptive Metadata, which gives information about the resource for identification such as title, abstract, author, and keywords.</p>
<b>OHDSI</b>	Observational Health Data Sciences and Informatics	The Observational Health Data Sciences and Informatics (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative to enhance the value of health data through large-scale analytics. OHDSI are the current owners and developers of the OMOP Common Data Model.
<b>OMOP CDM</b>	Observational Medical Outcomes Partnership - Common Data Model	<p>The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format. It enables the capture of information (e.g., encounters, patients, providers, diagnoses, therapeutics, measurements, and procedures) in the same way across different institutions.</p> <p>The purpose of a CDM is to standardise the format and content of observational data to apply standardised applications, tools, and methods across different datasets.</p> <p>Use of a CDM integrates medical records across healthcare organizations so that these data resources can be queried to answer important questions quickly and efficiently.</p>
<b>Pseudonymization</b>		It is defined within the GDPR as "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual".
<b>ScanReport</b>		The output file from a White Rabbit scan. It contains information on the tables, values, field types and data frequencies from a source data set (e.g., an MS Access DB, CSV file, SQL dB etc.). The ScanReport (also referred to as metadata extract) is sent to the HDR Technical Team to inform the OMOP mapping.



<b>SDE</b>	Secure Data Environment	Secure Data Environments were established by the NHS in England in 2024 and aims to ensure secure access across England to healthcare data for approved research projects led by academics, industry organisations or NHS employees such as clinical researchers. They have specific design features that allow approved users to access and analyse data without the data leaving the environment.
<b>SH</b>	Safe Haven	An alternate term for Trusted Research Environment (TRE). Terms may be used interchangeably.
<b>Structural Mapping</b>		The (often manual) process of mapping source data tables and fields to OMOP CDM tables and fields.
<b>SSH</b>	Secure Shell	A secure remote management protocol that allows network services to be operated over an unsecure connection.
<b>Term Mapping</b>		The process of mapping source fields values from one database to standard OMOP vocabulary.
<b>TRE</b>	Trusted Research Environment	Trusted Research Environments (TREs), also known as 'Data Safe Havens,' are highly secure spaces to be used by researchers accessing sensitive data.  They are based on the idea that researchers should access and use data within a single secure environment. In other words: the data resides in one secure location and researchers interrogate the data from its location. There is no data movement.  TREs have multiple layers of security and safeguards in place, designed to minimise the risk of data being misused.
<b>Virtual Machine</b>		A Virtual Machine (VM) is a computer resource that uses software instead of a physical computer to run programs and deploy applications (apps). One or more virtual "guest" machines run on a physical "host" machine. Each VM runs its own operating system and functions separately from the other VMs, even when they are all running on the same host. This means that, for example, a virtual MacOS VM can run on a physical PC.
<b>White Rabbit</b>		A Java tool developed by OHDSI to help prepare ETLs (Extraction, Transformation, Loading) of longitudinal healthcare databases into the OMOP Common Data Model (CDM). The main function of White Rabbit is to perform a scan of the source data, providing detailed information on the tables, fields, and values that appear in a database. This tool is used for structural mapping. White Rabbit is typically the first piece of software used in the ETL process.

 **Related articles**