CO-CONNECT

COVID - Curated and OpeN aNalysis and rEsearCh plaTform

Implementation Guide

03 October 2022

Welcome

CO-CONNECT are excited to have you as one of our trusted Data Partners.

This guide outlines the steps for Data Partners to facilitate their data sets to be discoverable via the CO-CONNECT platform. This includes the various stages of data governance, data hosting and mapping to a common standard (Observational Medical Outcomes Partnership, or OMOP) for querying.

The CO-CONNECT Team are a team of highly skilled experts here to support all Data Partners through each step. Each Data Partner will be introduced to key contacts within the team to guide them through the onboarding process.

It is recommended that the reader familiarises themselves with the Key Terms used throughout the document before continuing.

It is recommended that the reader watch our CO-CONNECT Explainer, Finding Data, and our Accessing and Analysing Data videos before continuing.

General enquiries can be emailed to *co-connect@nottingham.ac.uk*

Table of Contents

Supplementary Material

Note: Appendices not available by URL below will be sent alongside this document to Data Partners

Video	What is CO-CONNECT?
Video	Finding Data
Video	Accessing and Analysing Data
Appendix 1	Metadata Profile Registration Process for HDR Innovation Gateway
Appendix 2	Technical Documentation
Appendix 3	CaRROT-CDM
Appendix 4	Data Anonymisation
Appendix 5	Meta Data Standardisation
Appendix 6	WhiteRabbit
Appendix 7	Data Protection

Summary of Key Points

CO-CONNECT is a collaborative, Medical Research Council & Department of Health and Social Care-funded project, which allows researchers to assess the feasibility of requesting access to data held on cohorts across a number of different datasets. It aims to solve some of the issues with gaining access to data sets, which can be a difficult and long process.

Compounding this problem, it is often difficult to know whether a dataset will be useful in a particular project without getting access to it. CO-CONNECT aims to solve this issue by using utilising a metadata profile to build a set of rules to map the source database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Once applied to the source data by the Data Controller (DC), the data is standardised, allowing researchers to build discovery queries which are collected by the DC and run on the data set to assess its usefulness. This is split into two parts:

1. OMOP Mapping Activity

This activity is to prepare the dataset for use in Cohort Discovery. Data Contributors (DCs) will send the metadata profile to the CO-CONNECT team who will create the code to format the data in the OMOP format. The CO-CONNECT team will send this code to the DC who will apply this to their database and upload the data onto their own servers.

2. Query

Researchers will create discovery queries (e.g. males over 30 with asthma or female ex-smokers with diabetes). The DC, via the software, will pick up and run this query in their own database and send the results back to the HDR UK Cohort Discovery Search tool. These queries will be discovery only in nature, this means that the answers will only describe the data set and will not contain any Personally Identifiable Information (PII) and may not be used for any purpose other than discovery.

What data is being shared from Data Contributors to HDR UK?

No personally identifiable information will be shared during this process. Metadata will be shared including information on the fields, columns, tables, rows, and structure of the database. The actual dataset will not be shared with HDR UK. All queries will be run by software under the control of the DC and return results that do not contain any identifiable information, and only summary statistics. CO-CONNECT and HDR UK will not be processing personal data on your behalf as no personal data will be shared.

Where will the data be held?

All personal data will stay hosted at your organisation. HDR UK will only be sent metadata and summary statistics as the results of queries.

CO-CONNECT staff from Dundee or Nottingham will work with your team to advise on the format in which the metadata should be shared and based on this design, a ruleset to transform the original dataset. The transformation process will be carried out by your data engineers and not by CO-CONNECT. The dataset will not be sent to HDR UK or CO-CONNECT team, nor will we work directly with it or access it in any other way.

Who retains control of the data?

The dataset itself stays with the Data Contributor. HDR UK, CO-CONNECT nor any other party should be sent or gain access to the datasets through this process.

If a researcher wants to request access to the data set, they will be signposted to the relevant contact from the Data Contributor and be put through any processes or procedures as specified by the Data Contributor. HDR UK will not grant access to the data sets but might in some cases act as a referrer.

Vision Statement

Building a network to ensure COVID-19 immunity can be understood by linking leading data collections from across the United Kingdom.

Overview

Background

CO-CONNECT is a multi-million-pound research project to help scientists across the UK to easily access the data they need to understand COVID-19 and develop potential therapies and treatments (CO-CONNECT Video). It is funded by the Medical Research Council (Part of UKRI) and the Department of Health and Social Care (part of NIHR) in direct response to the pandemic.

The project is supported by the CO-CONNECT team based at the Universities of Dundee, Edinburgh, and Nottingham (lead institution). Each part of the team provides supporting capabilities in data handling, data curation to OMOP, data governance and data infrastructure. The CO-CONNECT data team can help and support the inclusion of your data; further details are provided below to inform your choice.

Purpose and Objectives

The current COVID-19 pandemic has caused millions of deaths, severely strained health systems and damaged economies across the world.

Data which can help us to answer these key research questions about COVID-19 have been collected across the UK by a range of research groups and within routine clinical primary and secondary care settings. The fragmented landscape of data collection and curation means that it can be challenging for public health groups and researchers to find and access the high-quality data they need at pace.

CO-CONNECT will:

- Provide a mechanism, via the HDR Innovation Gateway to allow researchers to find what relevant datasets reside where and under what access conditions.
- Create an infrastructure which enables trustworthy, fast, de-identified, secure analysis of data from across multiple organisations.
- Standardise COVID-19 Serology data collection across the UK.
- Align with the HDRUK National Core Studies programme.
- Answer key research questions about COVID-19 and the implications for patient outcomes.

Work Packages

WP1: Patient and Public Involvement (PPI) and Engagement

This WP will ensure the infrastructure is known, used, and sustained. The approach of CO-CONNECT is to go beyond simply a communication strategy. As well as the clear need for engagement into patient and public populations there is also a need to actively engage the research community who will utilise the services provided by the data flows and platform.

WP2: Standards

This WP will figure out which data fields need to be captured to answer the key research questions and then work with labs across the UK, LIMS providers and data sources to change processes and systems to capture this additional data.

WP3: Automated Pipelines

This WP is not building new infrastructure or a new platform, we are leveraging and connecting different existing components of infrastructure and automating the data flows between them. This automated solution will reduce the data access costs for each research project utilising the system and provide a legacy infrastructure which can be utilised beyond the duration of CO-CONNECT.

WP4: Data Curation

This WP will support Data Partners to on-board relevant subsets of the data partners' data into the infrastructure. This data will be standardised using the OMOP format. The approach of CO-CONNECT is to provide timely, automated discovery querying across data sets without requiring extraction all the data into a centralised location. This support is provided with no expectation or need of having access to the data held by the Data Partner.

Key Stakeholders

Data Partner – a body who has been funded and contracted through CO-CONNECT due to having data of relevance for the CO-CONNECT project

Health Data Research UK (HDR-UK) – a national institute whose mission is to unite the UK's health and care data to enable discoveries that improve people's lives. CO-CONNECT are working in partnership with HDR-UK to provide scalable and robust data infrastructure and services.

HDR Innovation Gateway – the portal where researchers can search, discover, and request access to datasets, tools, and resources for research purposes.

BC|Platforms – a global, modular, highly configurable platform for federated healthcare data with a proven record of accomplishment of delivering automated data harmonization solutions (case studies), operating out of the UK.

Public Benefits

Analysis of COVID-19 data has underpinned the UK's response to the global pandemic and informed public health policies accordingly. However, the data is collected and stored in multiple different organisations and institutes across the UK. This fragmented landscape presents challenges for public health groups and researchers to find and access the high-quality data they need at pace. It can take many months for research groups to contact organisations/research groups to understand the nature of the data they hold, complete data privacy impact assessments, obtain data governance approvals, agree data sharing and data management protocols, perform data indexing and linkage, and standardise and clean the data from multi-sources before research analysis can take place.

CO-CONNECT is a UK wide infrastructure aiming to provide a secure mechanism for government policy makers, public health analysts and researchers to discover which data is available from multiple sources across the UK and perform instantaneous high-level meta-analysis on that data.

This will:

- Enable the community to answer high-level questions in the public interest which can help the UK combat the pandemic much faster than without this infrastructure.
- Help government policy makers, public health analysts and researchers to see what data has already been collected to answer their specific research questions, reducing duplication.
- Reduce the burden on the taxpayer to fund an initiative to collect new data when existing data can be re-purposed to address a similar research question.
- Reduce the effort on each separate research group by virtue of automated linkage and data standardisation, reducing the time to undertake research and analysis urgently needed to save lives and help the UK resume normality.
- Facilitate the extraction of subsets of data from multiple data sources into a single Trusted Research Environment/Safe Haven (following additional project specific governance applications). This will help researchers carry out accurate investigations removing the need for meta-analysis, since analysis can be carried out on single data sets and then brought together in an existing TRE.

With the roll out of COVID-19 vaccines, such data and underpinning infrastructure is key to understand and monitor many issues such as the emergence of virus variants, length of immunity provided by vaccines and the health impacts of long COVID.

The key initial public benefit and focus for CO-CONNECT is to implement the Discovery aspect of the project. This will result in curated COVID data sets being queryable by researchers to determine their usefulness for their own research. Beyond this core feature, we have extra features that we will perform Research and Development on.

These extra features are opt-in per Data Partner and are covered in detail later in this document:

- Patient overlap detection – detect when the same patient appears in multiple data sets
- Meta-analysis queries – provides some quantitative analysis functionality for data partners who have opted in
- Extraction to TRE – semi-automated method of extracting query data (subject to further IG approvals) to a TRE

Patient User Group

The CO-CONNECT Patient User Group are an inclusive range of public members from across the UK. They are involved throughout the project, including:

- attending a variety of project meetings covering the four work packages;
- actively contributing to discussions;
- asking questions;
- developing our website; and
- raising issues that are of interest to the public.

Through this collaboration we make sure CO-CONNECT is grounded in the real world and creating a platform that has public benefit.

> *"COVID-19 has had a major impact on all our lives, and in many ways. It's essential for us to learn more about COVID-19 to be able to move forward and sharing healthcare data can help find the answers we need. Making sure people's healthcare data is kept safe was important for me and I find it reassuring to see how the team have made sure that it is protected. Working alongside the research team has been interesting and I feel like my questions and comments are taken on-board. I feel like I am making a difference." Antony Chuter, Patient and Public Involvement Lead*

Architecture Summary

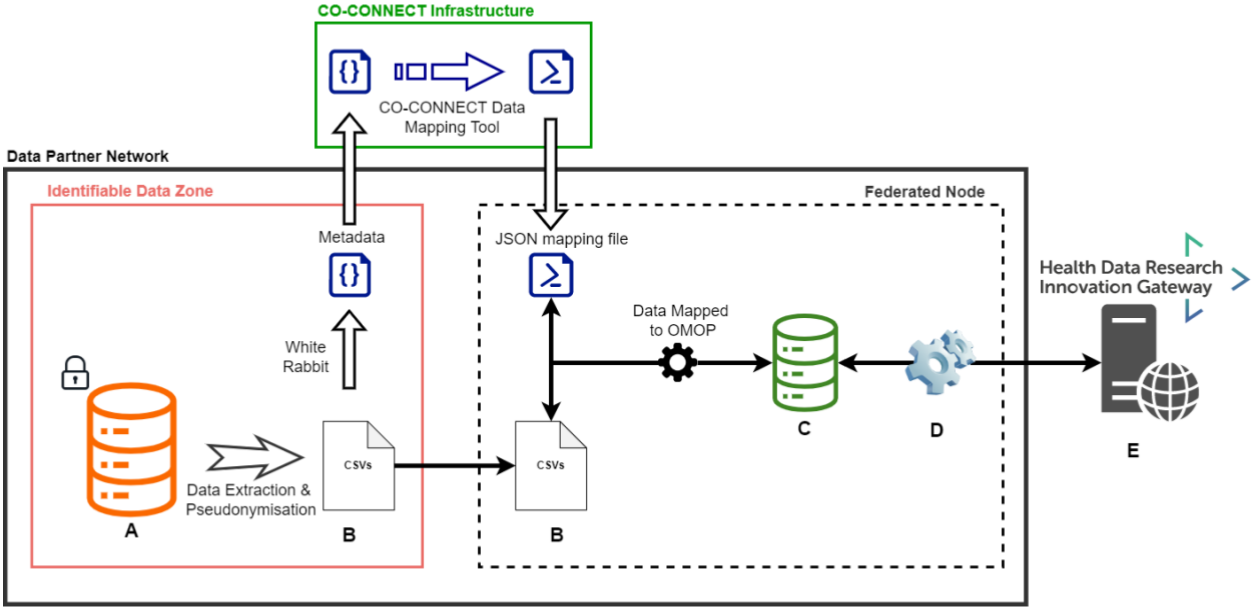

CO-CONNECT Infrastructure
CO-CONNECT Data Mapping Tool
Data Partner Network
Identifiable Data Zone
Federated Node
Metadata
JSON mapping file
White Rabbit
Data Mapped to OMOP
Health Data Research Innovation Gateway
Data Extraction & Pseudonymisation
CSVs
CSVs
A
B
B
C
D
E

Figure 1. Platform Architecture. The CO-CONNECT federated architecture. Each Data Partner (dark box) has potentially identifiable data (Dataset A) from which an extraction is made and pseudonymised (File B). A metadata extraction is performed with WhiteRabbit (within the Identifiable Data Zone, red box) and sent to the CO-CONNECT infrastructure (green box). A mapping script to the OMOP CDM is created using the CO-CONNECT data mapping tool (CaRROT-Mapper). The pseudonymised data are securely transferred (B) into a secure VM hosted by the data partner (Federated Node, dashed dark box), mapped to OMOP (CaRROT-CDM) and connected to federation software. From there the data are queryable by the Innovation Gateway(E)). Only aggregated, fully anonymous data discovery and meta-analysis is returned to the Gateway.

A secure virtual machine (VM) 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. Each Data Partner creates a database of relevant linked and pseudonymised datasets, in OMOP format (as detailed below), which are then inserted into the database hosted within their VM. A BC|LINK server is installed within the VM and connected to the pseudo-anonymised, OMOP standardised database. These localised BC|LINK instances are configured to communicate with a centralised tool within the HDR Innovation Gateway called BC|RQUEST.

Through the HDR Innovation Gateway, public health analysts, academic researchers and approved industry researchers can submit queries to BC|RQUEST. Locally, the BC|LINK within the VM requests these queries and runs them on the database, returning aggregate counts to BC|RQUEST and hence the user who initiated the query. This is simultaneously repeated across all the UK-wide Data Partners via the CO-CONNECT platform and all the results are displayed for analysis within the HDR Innovation Gateway. This functionality enables users to perform feasibility analysis (to discover relevant data from different sources) and carry out aggregate level analysis across different UK Data Partners.

The fundamental concept of CO-CONNECT is that Data Partners retain complete control of their data, whilst exposing metadata to allow researchers to investigate which cohorts exist and what data/information these cohorts are collecting across all Data Partners participating in CO-CONNECT. By using software from BC|Platforms, study feasibility questions can be asked, whilst all data remains within the Data Partner's secure environment and without any need to change security (i.e. inbound firewalls).

Following project specific data governance approvals, CO-CONNECT is researching methods that would allow the architecture to support semi-automated extraction of anonymised 'row-level' data from different Data Partners for analysis within an existing Trusted Research Environment (TRE) or Safe Haven (SH) (automated and streamlined). This step would require the research to have specific approval from each Data Partner and each Data Partner would agree which TRE should host all the datasets.

Governance Summary

CO-CONNECT has been designed from the ground up to allow Data Partners to retain full control of their data and to simplify the data governance requirements around their participation in this project. Mechanisms are in place to protect patient confidentiality and to ensure the security of the data. For example, identifiable data is only handled by the Data Partner and only within their own secure environment. The Data Partner will also pseudonymise their data before it is transferred to the secure VM within their own environment, and they have full control of which aspects of their data set CO-CONNECT can use in the project.

CO-CONNECT has been carefully designed to minimise the risk of someone running repeated queries on the system to re-identify a particular individual. Firstly, all Personally Identifiable Information (PII) (i.e. patient location, DOB, names etc) relating subjects in the data sets is to be withheld; no PII will be held in any software supplied by CO-CONNECT and managed by the Data Partner.

Secondly, only aggregated results for any discovery query will leave a Data Partner's control, subject to disclosure controls. Data Partners can control the low count suppression, for example counts of less than 5 are returned as 0 and counts between 5 and 10 can be rounded off to nearest 10 etc.. Furthermore, queries can only be constructed from pre-defined fields and users can only query data that has been authorised via a drag 'n' drop interface. Data Partners set both the threshold at which no results are returned and whether rounding should be applied.

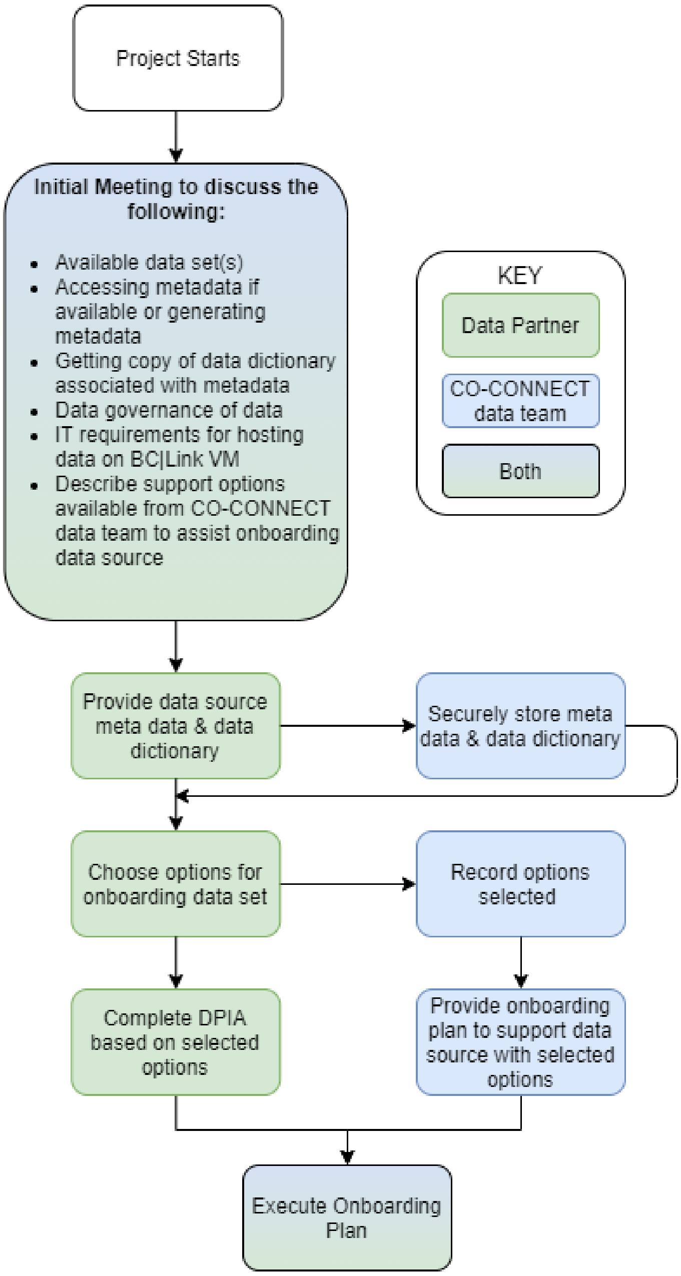
Data Partners are requested to carry out a Data Protection Impact Assessment (DPIA) although it is expected that as the data is sufficiently anonymised, the overall risk will be low.

Project Participation Steps

Step 1 Project Initiation

Figure 2 shows the initial steps required for initiating Data Partner's involvement in the project.

An initial joint meeting with the CO-CONNECT team is required. This will help to define the exact tasks and the responsibilities from that point onwards. The purpose of the meeting is to develop the plan for onboarding the data under the Data Partner's custody.


Project Starts
Initial Meeting to discuss the following:
Available data set(s)
Accessing metadata if available or generating metadata
Getting copy of data dictionary associated with metadata
Data governance of data
IT requirements for hosting data on BC|Link VM
Describe support options available from CO-CONNECT data team to assist onboarding data source
KEY
Data Partner
CO-CONNECT data team
Both
Provide data source meta data & data dictionary
Securely store meta data & data dictionary
Choose options for onboarding data set
Record options selected
Complete DPIA based on selected options
Provide onboarding plan to support data source with selected options
Execute Onboarding Plan

Step 2 Metadata Profile Registration Process for HDR Innovation Gateway

The HDR-UK Gateway includes a metadata catalogue from across multiple datasets and enables them to be searchable and accessible. Depositing metadata information in the Gateway will enable researchers and users to understand the data available. The catalogue will not contain any source data.

The metadata catalogue will help the researchers and innovators by providing:

- quick access to the meta-data 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 schema

To enable this functionality, the CO-CONNECT team will require Data Partners to prepare dataset metadata. (Appendix 1 – Metadata Profile Registration Process for HDR Innovation Gateway)

Step 3 Data Partner Implementation

A Data Partner's involvement in CO-CONNECT has multiple strands covering Information Governance (IG), Infrastructure and Data. Many of the activities involved in these strands can run in parallel. The IG processes should be started as soon as the project is initiated. The data, infrastructure and IG aspects of the project can all run in parallel.

Figure 3 highlights the infrastructure (I1-I3) and Data (D1-D6) steps involved in a Data Partner's involvement in this project.

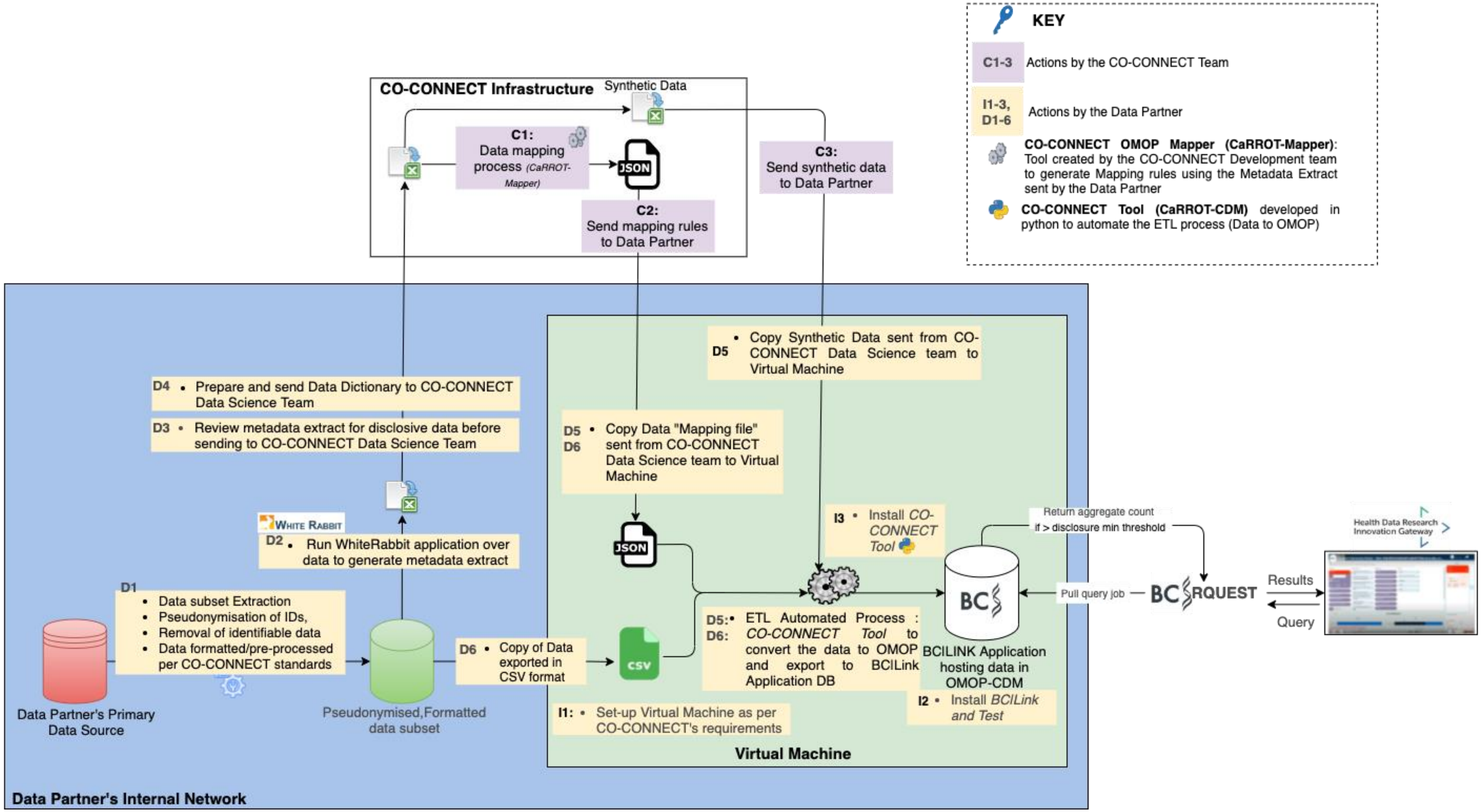

KEY
C1-3
Actions by the CO-CONNECT Team
I1-3, D1-6
Actions by the Data Partner
CO-CONNECT OMOP Mapper (CaRROT-Mapper): Tool created by the CO-CONNECT Development team to generate Mapping rules using the Metadata Extract sent by the Data Partner
CO-CONNECT Tool (CaRROT-CDM) developed in python to automate the ETL process (Data to OMOP)
CO-CONNECT Infrastructure
Synthetic Data
C1: Data mapping process (CaRROT-Mapper)
JSON
C2: Send mapping rules to Data Partner
C3: Send synthetic data to Data Partner
D5 • Copy Synthetic Data sent from CO-CONNECT Data Science team to Virtual Machine
D4 • Prepare and send Data Dictionary to CO-CONNECT Data Science Team
D3 • Review metadata extract for disclosive data before sending to CO-CONNECT Data Science Team
D5 D6 • Copy Data "Mapping file" sent from CO-CONNECT Data Science team to Virtual Machine
White Rabbit
D2 • Run WhiteRabbit application over data to generate metadata extract
I3 • Install CO-CONNECT Tool
Return aggregate count
if > disclosure min threshold
Health Data Research Innovation Gateway
JSON
D1
• Data subset Extraction
• Pseudonymisation of IDs,
• Removal of identifiable data
• Data formatted/pre-processed per CO-CONNECT standards
D6 • Copy of Data exported in CSV format
CSV
D5:• D6: ETL Automated Process : CO-CONNECT Tool to convert the data to OMOP and export to BCILink Application DB
BC
BCILINK Application hosting data in OMOP-CDM
Pull query job
BC RQUEST
Results
Query
Data Partner's Primary Data Source
Pseudonymised,Formatted data subset
I1: • Set-up Virtual Machine as per CO-CONNECT's requirements
I2 • Install BCILink and Test
Virtual Machine
Data Partner's Internal Network

Figure 2 - CO-CONNECT Standard Data Flow Model (CO-CONNECT Data Pipeline Videos)

Infrastructure Implementation Steps

I1: Setting up Virtual Machine

Responsible for this action: Data Partner

Who can see the data/process: Data Partner

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

Description: This virtual machine (VM) will provide a standard secure environment within each Data Partner's network to house and process the Data Partner's pseudo-anonymised data. On the virtual machine will reside the CO-CONNECT CaRROT-CDM Tool and BC|LINK application to perform data conversion to OMOP CDM and allow queries to be performed respectively.

Virtual Machine Technical specifications: 1 x Virtual Machine (LINK VM) to host link software with the following:

Operating System:

- CentOS or RHEL 7.9 minimum

Minimum/Recommended hardware:

- 4/8 vCPU
- 16/32 GB RAM
- 200/500 GB storage

I2: Install BC|LINK and Test

Responsible for this action: Data Partner assisted by CO-CONNECT infrastructure Team

Who can see the data/process: Data Partner

Where does the process take place: Within Data Partner's secure network environment.

Description: BC|LINK is an application developed by BC|Platforms which runs on a Postgres database. This application will be installed on the VM (explained in the previous section) to host the pseudo-anonymised data converted to OMOP CDM.

Each Data Partner, with the remote assistance of the CO-CONNECT infrastructure team (Appendix 2 – Technical Documentation), will be asked to configure and install the BC|LINK software application.

Once the BC|LINK application is installed, it will only communicate with the Query Portal (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.

I3: Software Installation for the automated ETL (Data to OMOP) Process.

Responsible for this action: Data Partner assisted by CO-CONNECT infrastructure Team

Who can see the data/process: Data Partner

Where does the process take place: Within a Data Partner's virtual machine in their own secure network environment

Description: Each Data Partner, with the remote assistance of the CO-CONNECT team if requested will install the CO-CONNECT CaRROT-CDM Tool to Extract, Transform and Load (ETL) tool (Appendix 3 – ETL) the Data Partner's data. This tool will apply a set of transformation rules generated by the CO-CONNECT Data Team, based on your source data's metadata, to the pseudo-anonymised source data, thereby transforming it to the OMOP format so it can be loaded into BC|LINK application.

Data Processing Implementation Steps

D1: Data Preparation

Responsible for this action: Data Partner

Who can see the data/process: Data Partner

Where does the process take place: Within Data Partner's environment (Secure Network)

Description: The steps involved in this action are stated below

A. Data Extraction

Data Partner will extract the relevant data for CO-CONNECT (i.e. Data Partner key health datasets) into flat files (CSV).

B. Data Pseudo-anonymisation

To ensure that the data is made discoverable without increasing the risk of re-identifying patients, following actions will be performed by the Data Partner:

1. CHI/NHS Numbers, and any other identifiers, will be pseudonymised using the method described in Appendix 4 – Data Anonymisation.
2. Patient identifiable information such as name, address, postcode etc. removed by the Data Partner.
3. Potentially identifiable information such as date of birth to be obfuscated by only providing the year of birth and replacing date and month with the 1st of January (if an existing date of birth obfuscation method exists then this can be used instead).

C. Data Standardisation

The CO-CONNECT Data team do not access the source data at any point in the process. To transform the source data to OMOP using the CO-CONNECT CaRROT-CDM Tool, it must be provided in a specific format. This ensures that the CO-CONNECT tools work effectively across all cohorts in a uniform manner. Please see Appendix 5 - Meta Data Standardisation and Data Standardisation Video Demonstration .

D2: Run WhiteRabbit application over pseudo-anonymised and standardised data to generate metadata extract

Responsible for this action: Data Partner

Who can see the data/process: Data Partner

Where does the process take place: Within a secure area of the Data Partners network that has access to the pseudonymised and standardised version of the data

Description: The CO-CONNECT Data team requires metadata in a standard format to start the data transformation process. Data Partners will be asked to download, install and run an open-source application called White Rabbit (Appendix 6 - WhiteRabbit , Metadata Extract Video Demonstration) to extract metadata from the source data. This application does not require an internet connection to operate, can connect directly to a database or delimited files, and is widely used by healthcare institutions globally. WhiteRabbit extracts aggregated metadata from the files or database, such as the table names, field names, and a frequency distribution of the values within those fields. The output document is called a ScanReport.

WhiteRabbit features a “minimum cell count” option; terms that exist within the dataset but appear with a frequency under this value will be automatically removed from the resulting ScanReport. For instance, if there are 3 patients in a dataset who have a rare disease, when the threshold is set to 5 the value will be automatically removed from WhiteRabbit output. The default threshold value is set to 5, but Data Partners can easily change this value to meet their security/governance requirements.

D3: Review Metadata Extract (WhiteRabbit Output)

Responsible for this action: Data Partner

Who can see the data/process: Data Partner

Where does the process take place: Within a Data Partner’s secure network environment

Description: Although the WhiteRabbit metadata extract (Scan Report) does not include any record level data, the output must be reviewed by the Data Partner to ensure that any information that the Data Partner considers sensitive is removed. In case there is any information that is considered sensitive then the metadata extract file must be manually removed by the Data Partner before sending it to the CO-CONNECT data team (Metadata Review Video Demonstration).

The WhiteRabbit metadata extract file will be e-mailed by the Data Partner as a password protected zipped file to the CO-CONNECT Data Team, with the password shared in a separate email.

D4: Provision of Data Dictionary

Responsible for this action: Data Partner

Who can see the data/process: Data Partner

Where does the process take place: Within a Data Partner’s network.

Description: The CO-CONNECT Data team will require a data dictionary containing table, column, and column value descriptions. It is the responsibility of the Data Partner to provide this information since it cannot be derived from the data itself. The format of the data dictionary can also be found on Appendix 5 - Meta Data Standardisation .

C1: Data Mapping Process

Responsible for this action: CO-CONNECT Data Team

Who can see the data/process: CO-CONNECT Team

Where does the process take place: Within CO-CONNECT's secure network environment.

Description: The CO-CONNECT Data team will use the WhiteRabbit ScanReport to generate a set of "mapping rules". These rules contain guidelines for the ETL conversion to OMOP.

C2: Send Mapping Rules to Data Partner

Responsible for this action: CO-CONNECT Data Team

Who can see the data/process: CO-CONNECT Team

Where does the process take place: Within CO-CONNECT's secure network environment.

Description: This "mapping file" is in JSON format and will be provided to the Data Partner via e-mail as it does not contain any sensitive information. The process of generating this file can be found on Generating "Mapping rules" file Video Demonstration.

C3: Send Synthetic Data to Data Partner

Responsible for this action: CO-CONNECT Data Team

Who can see the data/process: CO-CONNECT Team

Where does the process take place: Within CO-CONNECT's secure network environment.

Description: The CO-CONNECT Data team will generate synthetic data based on the WhiteRabbit ScanReport received from the Data Partner. This synthetic data is artificially created data which reflects the properties of the real data such as table names, field names and field values. This synthetic data will be provided to the Data Partner (in csv format) to allow them to test the CO-CONNECT CaRROT-CDM ETL Pipeline with data that is representative of their own.

D5: Virtual Machine and Software Testing using Synthetic Data and Mapping File

Responsible for this action: Data Partner assisted by CO-CONNECT Infrastructure Team

Who can see the data/process: Data Partner

Where does the process take place: Within Data Partner's secure network environment.

Description: Once BC|LINK and the CO-CONNECT CaRROT-CDM Tool are installed and configured, the CO-CONNECT team will assist the Data Partner to test the Virtual Machine (VM), BC|LINK application and the CO-CONNECT CaRROT-CDM ETL tool (Appendix 3– ETL).

A. Copy Synthetic Data and Mapping Files to the Virtual Machine

The Data Partner will receive synthetic data reflective of their data and the "mapping file" from the CO-CONNECT Team. The Data Partner will import this synthetic data and the "mapping file" to the VM.

B. Run CO-CONNECT CaRROT-CDM Tool

Once the files are copied the automated CO-CONNECT ETL process will be triggered and the CO-CONNECT CaRROT-CDM Tool will:

1. Extract the synthetic data files (CSV) and the mapping file (JSON) from the supplied file location on the VM.
2. Transform the Data to OMOP-CDM.
3. Load the data into BC|LINK application.

Once the synthetic data is live on the BC|LINK, to validate the testing, the CO-CONNECT team will run specific test queries using the connected test instance of BC|RQUEST against the Data Partner's synthetic data confirm the aggregates meet expectations.

C. Send logfile to CO-CONNECT Data Team

After running the CO-CONNECT CaRROT-CDM ETL tool on synthetic data, a log file called "coconnect.log" will be generated which will contain details of the ETL process performed by the CO-CONNECT CaRROT-CDM ETL tool. As part of the testing process, this file will be provided to the CO-CONNECT team to confirm the tool runs as expected. The log file will not contain any sensitive information as it will generate logs based on the synthetic data. However, the Data Partner is still advised to view the log file before sending it to the CO-CONNECT team to make sure they are happy sharing its contents with us.

D. Network Security & Vulnerability Testing

The Data Partner's IT security team can monitor the network traffic of the BC|LINK and run vulnerability or penetration tests as required.

Once the VM and software has been tested, the CO-CONNECT team will guide the Data Partner through the process of altering the configuration to make the LINK application live.

The CO-CONNECT team will not have direct access to the VM at any point in this process.

D6: Run the automated ETL Process on Pseudo-anonymised Data

Responsible for this action: Data Partner

Who can see the data/process: Data Partner

Where does the process take place: Within a Virtual Machine set up in Data Partner's secure network environment.

Description: Once the secure VM is set up, tested, and verified by the CO-CONNECT team, the Data Partners will be required to copy their standardised and pseudo-anonymised data in CSV format and the "mapping file" to the Virtual Machine.

Once the files are copied, the automated ETL process will trigger the CO-CONNECT CaRROT-CDM Tool to extract the pseudo-anonymised data files (CSV) and the mapping file (JSON) from the file location, transforming the Data to OMOP-CDM and loading the data to BC|LINK application. Once the data is loaded to BC|LINK, it can be queried via the HDR Innovation Gateway Cohort Discovery Tool. For the Data Partners with regular data feeds, this process will iterate at regular intervals (weekly, monthly etc). This frequency will be agreed between CO-CONNECT and the Data Partner. This will ensure that new data is ready for external querying and discovery as it becomes available. (ETL Process Video Demonstration).

Running Data Discovery Queries

The BC|RQUEST application has been integrated into the HDR UK Gateway and is called the Cohort Discovery Tool. This application acts as a central query engine for users of the Innovation Gateway.

The Data Partner's BC|LINK application within the local VM will:

1. Check for query job(s) from BC|RQUEST (within the HDR Gateway) every 5 seconds
2. Run these queries on the underlying pseudonymised OMOP database and
3. Return aggregate level counts of the number of individuals within the dataset which meet the query criteria.

The communication between the BC|LINK software and the central query engine uses an encrypted SSH tunnel. This is an automated process using Public/Private Key Pair Authentication. Only authorised/approved users will be able to initiate queries via the HDR UK Innovation Gateway (see *Data Governance and Security Controls* section for full details how users are approved to gain access to run queries).

Step 4 Project Governance Implementation Steps

Responsible for this action: Data Partner

Who can support this process: CO-CONNECT

When does the process take place: In parallel with Steps 1-3

Description: Data Partners are likely to also be the Data Controllers. In the event that a Data Partner is not a Data Controller, consent must be obtained from relevant Data Controllers regarding this use of the data.

CO-CONNECT is built on a flexible technical philosophy that means Data Controllers can participate with their current governance and consent. Discovery and meta data analysis are performed without requiring the data to move.

Data Partners will need to carry out a Data Protection Impact Assessment (DPIA) to review potential risks to Data Governance and Security Controls and patient confidentiality with the use of the CO-CONNECT platform. Whilst it is expected that the DPIA assessment will conclude that participation is low risk and the data will be sufficiently anonymised that it is not in remit of GDPR, but it is important Data Partners make that conclusion independently. There is data governance support available if required.

Data Partners may require additional internal data access applications to be carried out, this will be the responsibility of the Data Partner although full support from the CO-CONNECT team can be provided.

A technical risk assessment may be required by your organisation; this is the responsibility of the Data Partner although full support from the CO-CONNECT Team can be provided.

Dependency: Step 3 only synthetic data can be used to test the system until Step 4 is complete.

Iterative On-boarding of Data Sets and Fields

CO-CONNECT supports an iterative approach for on-boarding a range of data cohorts and data fields.

Onboarding multiple data cohorts

A Data Partner with multiple data cohorts to be on-boarded into CO-CONNECT can deploy a single institutional BC|LINK instance/server since a single BC|LINK server can store and make multiple cohorts discoverable. However, additional configuration steps will be required by the CO-CONNECT infrastructure teams to enable this capability and hence Data Partners with multiple data cohorts are asked to contact the CO-CONNECT Team for further support.

Onboarding new data fields to an existing cohort

Within each of the datasets, the most important fields have already been identified by the CO-CONNECT team with the assistance of medical researchers associated with the project, SAGE and NCS (National Core Studies). This process has led to a set of key questions that needed to be answered and supported the CO-CONNECT Team to focus on data such as serology tests, vaccinations, ethnicity, multi-morbidities, prescribing data, death data and GP data.

If additional new data fields are identified, the process of including these additional data fields will be iterative, first focusing on the key fields required to answer pressing public policy, public health and research questions. This process will allow the addition of new fields requested by the community to be added over time.

The iterative on-boarding process for both new collections and/or data will follow the same governance process explained in section *Data Governance and Security Controls*.

Data Governance and Security Controls

Following on from Step 4 Project Governance Implementation Steps, Discovery and metadata analysis are performed **without requiring the data to move**.

CO-CONNECT have put several measures in place to protect and safeguard data (Appendix 7 – Data Protection).

As discussed in Step 4 Project Governance Implementation Steps, Data Partners should carry out a Data Protection Impact Assessment (DPIA). The CO-CONNECT Team can provide data governance support if required in the form of supporting documents, assistance filling in the DPIA or support answering any IG related questions. We find that a Q&A session with IG teams that includes a demo of the Cohort Discovery tool tends to be a good way to progress.

There are several key controls which will be in place to protect patient confidentiality and data security:

- Identifiable data is only ever handled by Data Partner employees within the Data Partner IT infrastructure.
- Data is pseudo-anonymised by the Data Partner before being copied to a secure VM.
- The pseudonymised record level data will not be available outside of the Data Partner's infrastructure.
- Firewalls will be in place with standard controls against malicious attacks and require no additional inbound rules along.
- The CO-CONNECT Team are highly experienced in handling consented and non-consented data.

Disclosure control on queries

- Only metadata and aggregated results for any query (discovery and meta-analysis) will leave the Data Partner's control. The Data Partner can control the low count suppression for e.g., counts of less than 5 are returned as 0 and counts between 5 and 10 are rounded off to nearest 10 etc. Once the test version of the BC|LINK is installed in the Data Partner's environment they can tweak these disclosure controls setting and test them out on the synthetic data we provide so they can find the configuration that best suits their organisation.
- Queries can only be constructed from pre-defined fields. A 'drag and drop' interface ensures that users can only query data that has been authorised.
- All queries are logged and can be reviewed by the Data Partner from the log management software.
- Data Partners will choose the scope and permissions to run meta-analysis queries on the data. For example, permissions could just be set to public health analysts rather than researchers. Requests to run meta-analysis will be managed and approved by the Data Partner.
- Researchers gaining access to aggregate data via the HDR Innovation Gateway must be authenticated users and associated with an appropriate organisation (this approval process is managed by HDR UK).
- Researchers gaining access to aggregate data via the HDR-UK Gateway will go through an approval process, ensuring they are associated with an appropriate organisation. The first is via OpenAthens.

- OpenAthens is a UK-wide identity federation that allows anyone with a recognised institution to login to any service under the federation. They use their institutional account to login and on sign up are required to agree to the HDR Gateway T&Cs. To access the CO-CONNECT Platform (Cohort Discovery in HDR terminology) they have a further set of T&Cs they need to review and accept before they can run queries.

  The second method is via logins using LinkedIn or Google. These users can request access to the CO-CONNECT Platform, but this redirects them to a ticketing system which requests extra information about them and an email verification check. This information is then manually verified to confirm they are a bona fide researcher before they are given access. HDR plan to review these types of accounts every six months to ensure they still have a valid reason to access the query portal.

  A person may receive bona fide researcher status if:

  1. Their home institution confirms they are a current researcher, OR
  2. A person who satisfies condition (1) corroborates their researcher status (as a reference)

  These are based on the GA4GH definition of a bona fide researcher as defined here.

Disclosure control on research projects

- Researchers seeking access to record-level data will need their own additional permissions from the Data Partner via their standard governance processes to undertake the research; the data will need to be stored in an accredited Safe Haven/Trusted Research Environment. This extraction process is a later phase of CO-CONNECT and will be Opt-in i.e., additional approvals will be required above those for aggregate level querying. More details of this feature will be covered in the next section.
- For each research project the Hashed CHI/NHS pseudonymised identifiers will be replaced by project specific unique identifiers unique to avoid the possibility of data linkage across multiple research projects.

CO-CONNECT Extended Features in the R&D Programme

When the Discovery aspects of CO-CONNECT have reached significant adoption (with the assistance of BC|Platforms), three new functionalities will be researched and added to support the research community to enhance their experience of cohort discovery and data access. All options are opt-in and would only activate with explicit permission from the Data Partner.

- **Patient Overlap Detection:** The Patient Overlap Detection functionality will enable a query run on the Discovery portal to identify and report duplicate records across different cohorts. This new functionality increases the precision of the results and also helps to identify cohorts that are recording the same data for a particular participant/patient. Thus, if two or more cohorts have the information the researcher is interested in, he/she will be able to identify the degree of record overlap between these cohorts.

- **Meta-analysis Queries:** The Discovery portal is limited to qualitative analysis only; the Meta-analysis Queries feature will provide quantitative analysis functionality, offering a variety of statistical analysis tools for Data Partners who have opted-in for this feature.

- **Data Extraction to TRE:** The Data Extraction to TRE feature will support semi-automated extraction of a subset of linked and pseudo-anonymised record level data from a Data Partner's BC|LINK. Any researcher who is interested in row level data will be required to go through the project specific governance approvals and will only have access to the approved project specific subset of the data.

Additional Per Query Pseudonymisation

For each query, the pseudonymised NHS/CHI identifiers will be replaced by a specific encrypted one-time ID for each person so in effect the IDs are double hashed for each query. For example, participant A with pseudonymised identifier 1234 is replaced with query-level identifier 001 in all cohorts (Appendix 4 – Data Anonymisation). This:

- Provides an additional layer of security on top of the hashed CHI/NHS number.
- Enables the identification of patient overlap in results (individuals which appear in multiple cohorts because of the query) which is a later phase of CO-CONNECT and is covered in the Extended Features section.
- Enables linkage of data from different data sources for data analysis of row-level data within a Safe Haven/Trusted Research Environment subject to appropriate permissions.

Patient Overlap Detection

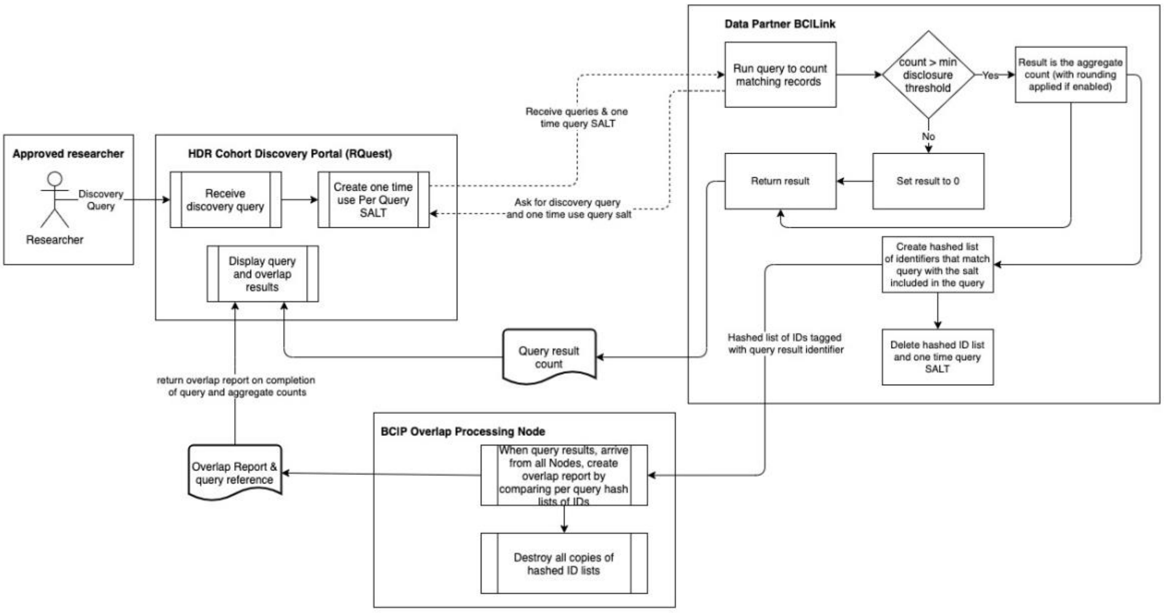

Approved researcher
Researcher
Discovery Query
HDR Cohort Discovery Portal (RQuest)
Receive discovery query
Create one time use Per Query SALT
Display query and overlap results
Receive queries & one time query SALT
Ask for discovery query and one time use query salt
Data Partner BCILink
Run query to count matching records
count > min disclosure threshold
Yes
No
Result is the aggregate count (with rounding applied if enabled)
Return result
Set result to 0
Create hashed list of identifiers that match query with the salt included in the query
Delete hashed ID list and one time query SALT
Hashed list of IDs tagged with query result identifier
Query result count
return overlap report on completion of query and aggregate counts
BCIP Overlap Processing Node
When query results, arrive from all Nodes, create overlap report by comparing per query hash lists of IDs
Destroy all copies of hashed ID lists
Overlap Report & query reference

Figure 3- Patient Overlap Process

The aim of Patient Overlap Detection function is for researchers using the HDR Cohort Discovery Portal to understand any overlap in the patients returned from multiple data collections after running a query on the Discovery portal. For instance, if 250 patients from dataset A meet the query criteria, and 140 patients from dataset B meet the criteria, but 130 of the patients in dataset B are also present in dataset A, the researcher knows that undertaking a study on both datasets is not necessary, since most patients exist in dataset A.

The first stage of Patient Overlap Detection process is to pseudonymise the NHS ID/CHI using a shared CO-CONNECT SALT which is applied to each data collection (**Error! Reference source not found.**). The second stage of the process involves the HDR Cohort Discovery Tool and creates a 'one-time use' SALT that is included with each discovery query submitted. This protocol will be received, and used by each connected BC|LINK.

Example: Table 1 - Query Aggregate Results

Dataset	Num Patients
A	15,500
B	260
C	21

When the BC|LINK receives the query with the one-time-use SALT, it runs the query and produces the aggregate counts for matching records with disclosure controls applied (Error! Reference source not found.Error! Reference source not found. minimum suppression value). The BC|LINK will then produce a list of the pseudonymised IDs that match the query and then apply the one-time-user per query SALT to each ID.

Example: Table 2 - Dataset Overlap Report

Dataset	A	B	C
A	15,500	116	0 [under disclosure threshold]
B	116	260	0
C	0 [under disclosure threshold]	0	21

This hashed list of identifiers will be sent to the overlap processing node with a result identifier. Once all BC|LINK Partner nodes have returned results, the processing of this node will compare the hashed list of identifiers to determine which records are shared between collections. The results will be provided in an overlap report that is sent back the to the Cohort Discovery Tool to be displayed alongside the query results to indicate how many records returned are unique (Table 2 is an example of an overlap report).

Once used, all hashed ID lists are destroyed (along with the one-time use SALT) from all nodes. The HDR Cohort Discovery Tool removes and destroys the one-time use SALT from the query, before archiving the result for future use.

It is important to note that the patient overlap results will be subject to small number suppression; this means that each Data Partner can set a threshold value, and patient overlap numbers below this value will not be shown to researchers.

Meta-Analysis Queries

The BC|Platforms architecture has the capability to enable meta-analysis queries on standardised, pseudonymised data. For example, a meta-analysis query can analyse quantitative antibody level results over time (potentially broken down by ethnicity via a trendline produced using the BC|Platforms tool). This example is the Minimum Viable Product (MVP) that the wider CO-CONNECT project wants to achieve.

The meta-analysis capability will only ever return aggregate level results and will also be subject to agreed low number suppression control.

Each Data Partner will be given an admin account for the query portal, hosted within HDR, to allow configuration of access controls to control their data sets. More specifically, the Data Partner will control the scope of queries, and access permissions provided to the researchers based on their governance approvals. The BC|LINK application allows the Data Partner to turn off the meta-analysis feature if they wish to opt out of this functionality on the HDR Innovation Gateway.

Semi-automated extraction to a TRE

Researchers will be able to run queries and meta-analyses (if this feature is enabled by the Data Partner) via BC|RQUEST from the BC|LINK application. These queries will return an aggregated result. However, if the researcher is interested in a row-level data subset of the query, this can be facilitated by the Data Partners using the BC|LINK application on the VM. The subset can be provided in the Data Partner's secure network environment for review before making the data available on a Trusted Research Environment for the researcher.

This functionality is only available as an opt-in feature for Data Partners. Researchers interested in the row-level data will have to go through required data governance procedures set by the Data Partner and will only have access to the approved subset of data.

This process can only be run by a trusted administrator, appointed by the Data Partner, who is a member of their organisation and helps to manage and control access to their BC|LINK. The process will allow the administrator to locate the query the researcher had run and used for their governance application. The BC|LINK can then use this query to extract the row level data from the Data Partner's BC|LINK and extract it into a secure zip file that the administrator can download to their secure environment. From this point, they can decide how and where to share the row level data with the researcher.

Contracts

Each Data Partner will be asked to sign a collaboration agreement with University of Nottingham as CO-CONNECT's Lead Institution. Each Data Partner will be asked to sign a BC|Platforms license agreement for the use of their software.

Key Terms

Term	Stands For	Description
BC|LINK		BC|LINK Application (App) is a piece of software which is sat within the Data Partner's infrastructure (Virtual Machine). This application runs on PostgreSQL (a relational database management system) which hosts the Data Partner's pseudonymised data. The App receives query requests from the query portal and translates it to SQL which is then run against the pseudonymised OMOP database. Summary statistics are returned to the end user at the query portal on the HDR Innovation Gateway. BC|LINK is developed by BC|Platforms. 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		It is a global leader in providing a powerful data and technology platform for personalized medicine and drug development, accelerating the translation of insights into clinical practice.
CO-CONNECT	Curated and Open aNalysis aNd rEsearCh plaTform	CO-CONNECT will support access to information from 44 data sources on behalf of a consortium of researchers, standardising antibody data collection from across the UK to build a secure and trustworthy federated platform for researchers to access. The system will also protect patient confidentiality and data security, supporting federated anonymised data analysis.

CaRROT-Mapper	CO-CONNECT OMOP Mapper	A web-tool designed and developed by the CO-CONNECT Software team that enables the CO-CONNECT 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).
CaRROT-CDM Tool		An ETL tool designed and developed by the CO-CONNECT Software team. This tool automates the Extraction of Pseudonymised data, Transformation of data to OMOP CDM and Loading to BC|LINK Process.
Data Controller		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. 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.
Data Curation		The organization and integration of data collected from various sources.
Data Custodian		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. 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.

Data Dictionary		Information about data such as table and field descriptions, relationships to other data, origin, usage, and format.
Data Discovery		The process of obtaining actionable information by finding patterns in data from multiple sources with interactive visual analysis. The term is used to express a mode of analysis in which researchers attempt to get a holistic view of all their data sources. The two main ingredients are the ability to join data sources and the interactive visual analysis component which allows exploration of the data to find patterns.
Data Flow		A dataflow is a path for data to move from one part of the information system to another. A dataflow may represent a single data element such the Customer ID or it can represent a set of data elements (or a data structure).
Data Governance		Managing data assets throughout their lifecycle to ensure they meet organisational quality and integrity standards.
Data Interoperability		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.
Data Partner		A body who has been funded and contracted as part of CO-CONNECT through their local institutions.
Data Processor		A natural or legal person, public authority, agency, or other body which processes personal data on behalf of a Data Controller.
Dataset		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.
DPIA	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. It is also good practice to do a DPIA for any other major project which requires the processing of personal data. The DPIA must describe the nature, scope, context, and purposes of the processing; assess necessity, proportionality, and compliance measures; identify and assess risks to individuals; and identify any additional measures to mitigate those risks.
ETL	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. In the context of the CO-CONNECT project, preparation for the ETL is supported by White Rabbit, Rabbit in a Hat and Usagi.
GDPR	General Data Protection Regulation	A legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU).
Hashing Algorithm		A mathematical algorithm that maps data of arbitrary size (often called the "message") to a bit array of a fixed size (the "hash value", "hash", or "message digest"). It is a one-way function, practically infeasible to invert. There are different hashing algorithms for e.g., MD5, SHA-2 etc.

HDR UK Innovation Gateway	HDRUK	A portal enabling researchers and innovators in academia, industry, and the NHS to search for and request access to UK health research data.
Mapping Database		A database to be developed by the CO-CONNECT Data Management Team to define how source data will be mapped to the OMOP CDM.
Meta Data Search		The ability to find a Data Source based on the data fields. As an example, finding all Data Sources that collect ethnicity and immune response level.
Meta-analysis		Meta-analysis is the statistical procedure for combining data from multiple studies. When the treatment effect (or effect size) is consistent from one study to the next, meta-analysis can be used to identify this common effect. When the effect varies from one study to the next, meta-analysis may be used to identify the reason for the variation.
Metadata		Metadata is information about the data and this document refers to two types. Structural Metadata, which gives information about the table names and field names in each table for each data set. Descriptive Metadata, which gives information about the resource for identification such as title, abstract, author, and keywords.
OHDSI	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.

OMOP CDM	Observational Medical Outcomes Partnership - Common Data Model	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. 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. 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.
Pre-Processing		A term used by the CO-CONNECT Data Team to describe any transformations that the Data Partner applies to the source data to ensure it conforms to the CO-CONNECT Data Standards.
Pseudonymisation		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”.
Record Level discovery		It is the ability to find a Data Source based on the values of a particular field. For example, how many people across the Data Sources have a positive immune response and are of Black, Asian, and Middle Eastern ethnicity.

Salt		A salt is extra data (random or otherwise) that is used as an additional input to a one-way function in order to make the hashed output which makes a data breach more difficult. Salt is an extra layer of security. Salts are used as a safeguarding measure.
ScanReport		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 is sent to the CO-CONNECT Data Team to inform the OMOP mapping. In the context of CO-CONNECT it is often referred to as metadata.
SH	Safe Haven	An alternate term for Trusted Research Environment (TRE). Terms may be used interchangeably.
Structural Mapping		The (often manual) process of mapping source data tables and fields to OMOP CDM tables and fields.
SSH	Secure Shell	A secure remote management protocol that allows network services to be operated over an unsecure connection.
Term Mapping		The process of mapping source fields values from one database to standard OMOP vocabulary.
TRE	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.
Virtual Machine		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.
White Rabbit		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.