CO-CONNECT Data Anonymisation

Preparing your data for discovery

November 2021

Introduction

This document provides information on the data anonymisation requirements relating to the CO-CONNECT project, referred to in CO-CONNECT Implementation Document, Data Processing Implementation Steps, Section D1 and in CO-CONNECT Extended Features. It will describe what those requirements are, why they are important, and what needs to be implemented to enable participation in the CO-CONNECT project.

Background

The CO-CONNECT project has been established to facilitate data discovery whilst always ensuring that confidentiality is maintained, and identifiable information is not exposed. As such, there are certain steps that need to be undertaken to ensure that the Personally Identifiable Information (PII) is either removed, encrypted, or obfuscated prior to interacting with CO-CONNECT. This is to ensure that from a software and wider CO-CONNECT project team perspective, the data can be considered anonymous.

The CO-CONNECT Project Team as well as the software it provides should never have access to data or information that could identify individuals[1].

In terms of scope, it is highly likely that Data Partners will be holding both confidential information and PII as part of their research cohorts and routinely collected data. Note this document only relates specifically to the management of data that could identify individuals – for example, patient level identifiers such as NHS numbers.

Additional steps like data aggregation will also be undertaken to ensure those using CO-CONNECT will never have access to individual record level data. These steps are documented elsewhere[2].

NHS & CHI Numbers

The 'linking key' or core identifier, NHS (and CHI numbers in Scotland) enable individuals to be linked across multiple datasets. From a research perspective, this is particularly powerful and as a result it is a core feature of CO-CONNECT. However, in their original formats NHS and CHI numbers could lead to the inadvertent patient identification.

As a result, before any data reaches software (i.e. BC|LINK[3]) provided by CO-CONNECT, each Data Partner must ensure that NHS or CHI numbers are processed through a pre-defined hashing algorithm – namely SHA-256[4].

NHS numbers can be recorded in multiple formats. For example, 123 456 7890, 123-456-7890, 123/456/7890, etc. To ensure consistency across datasets, each Data Partner must remove any non-numeric characters from their source identifiers before a salt or hashing is added.

For the previous examples, these would all resolve to 1234567890, meaning they can be linked together after modification. This process should also be performed for CHI numbers in Scottish datasets.

[1] The CO-CONNECT project teams are based in Nottingham, Dundee and Edinburgh.
[2] See document titled *CO-CONNECT & Data Protection*.
[3] BC|LINK is a tool co-located in the Data Partner's IT environment that enables aggregate level data to be queried by the CO-CONNECT user facing portal.
[4] See appendix for further information.

As per best practice, CO-CONNECT will provide a single salt for use across all Data Partners. Salts will be encrypted, securely distributed, and made accessible to participating Data Partners. The provision of a single salt is to ensure that data brought together can still be linked (if permissions allow), without having to exchange the raw identifier.

In a second phase of discovery work this will also allow the federated search to both report a total count, and a unique individuals count, as the overlap between studies can be discovered.

Study Identifiers

If a dataset contains a study-specific identifier rather than an NHS or CHI number, then the above considerations are not relevant. Instead, it is accepted that the identification of patients who exist in multiple datasets will not be possible.

Study identifiers can be included without any anonymisation, however if they are considered sensitive by the Data Partner, they can be secured according to the internal governance requirements of that Data Partner. For instance, a Data Partner may choose to hash the identifiers with their own salt, replace them with sequential numbers or any other anonymisation method.

Dates

For best practice, several additional steps should also be undertaken

1) Date of Birth should be set to year only, with month and day set to "01". For example, 1980-04-23 00:00:00 becomes 1980-01-01 00:00:00.
2) Event Dates which may be identifying need to be reviewed to decide whether adjusting dates to the first of the month would reduce the risk of identification.

CO-CONNECT & Salt Key Management

Salt keys will be distributed via OneDrive to a nominated individual (or individuals) within the participating Data Partner. This person(s) should be identified to the CO-CONNECT Project Team. NB OneDrive settings will be such that only the nominated individual(s) can access the respective file; access will also be tracked and monitored.

Salt Key Data breach

In the event salt keys are compromised – or indeed, are thought to be compromised – the following steps will be taken

1) Access to salt key files will be restricted and / or deleted by the CO-CONNECT Project Team
2) Data Partners will be asked to remove their existing data from BC|Link
3) A new salt key will be generated by the CO-CONNECT Project Team and distributed via OneDrive to all participating Data Partners
4) Data Partners will need to
    - Access the new salt key
    - Re-anonymise their data using the new salt key
    - Re-run the OMOP transformation script
    - Upload the newly anonymised data to BC|Link

5) An investigation will then be conducted in conjunction with the Data Partners to ascertain the source and cause of the breach
6) Mitigating actions will be carried out to manage and restore data security

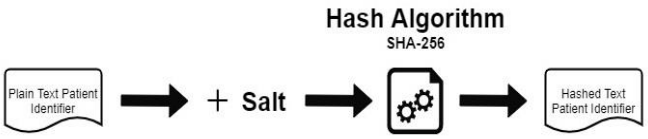
Link to Code Repository

To assist and to provide more information on SHA-256, Data Partners are invited to visit https://github.com/CO-CONNECT/Pseudonymisation.

Appendix

Secure Hash Algorithm – SHA 256 & Salt

SHA-256 is a one way / irreversible cryptographic hashing function approved by the US National Institute of Standards and Technology (NIST). In addition, the Information Commissioner's Office (ICO) approves the use of cryptographic hashing functions as being suitable for anonymising data[5]. SHA-256 and the wider SHA-2 family are also commonly used in a variety of sectors including healthcare and financial services.



[5] https://ico.org.uk/media/1061/anonymisation-code.pdf (p69). NB this is pre-GDPR; updated guidance from the ICO is expected soon.