Comprehensive Patient Records for Cancer Outcomes

Data Flow Protocol

Table of Contents

1. Purpose

This protocol sets out how, for the 'CPR for Cancer Outcomes' study:

- Data will flow from health records to a research environment and beyond
- OpenPseudonymiser[1] will be used to enable only authorised linkage

It provides a lay summary and a technical review.

2. Lay Summary

The purpose of the 'CPR for Cancer Outcomes' study is to **securely link information that is non-identifiable (not linked to named individuals) from electronic GP and community (primary care) and hospital records.** This information (known as data) will provide the means to get a very clear picture of what happens to cancer patients across their cancer pathway. This protocol explains how data linkage will be done.

The data will be linked from:

- Patient hospital records held at the Leeds Teaching Hospitals NHS Trust (LTHT)
- Non-identifiable primary care records that have been opted in to ResearchOne (R1), a research database held by The Phoenix Partnership (TPP). TPP is a clinical systems supplier to over 5,000 healthcare organisations in the UK.

Appendices 3 and 4 give further information about where the data comes from.

LTHT and TPP will provide data from records of a group (known as a cohort) of cancer patients and cohort of control patients. The data will be linked at the University of Leeds Integrated Research Campus (IRC). It will be linked using codes known as digests created using a safe and protected, well-tested process called OpenPseudonymiser. Data will remain securely in the IRC while it is used for research.

The following sub-sections explain OpenPseudonymiser and how data will be handled during the project.

2.1 OpenPseudonymiser

The 'OpenPseudonymisation process' enables data linkage without the need for identifiable information. Where datasets previously may have been linked on NHS numbers, OpenPseudonymiser turns the NHS numbers into codes known as digests that can be linked on instead.

OpenPseudonymiser also helps to control what datasets are linked. For example, previously any data with NHS numbers could be linked in at any point. In contrast,

[1] https://www.openpseudonymiser.org/

OpenPseudonymiser creates digests for datasets within a project by using a project-specific key. This means that the digest created from an NHS number for one project would not match the digest created from the same NHS number for a different project. As such, the digest could not be used to link data from different projects. This helps to prevent unauthorised or unethical linkage.

2.2 Data Flow

Non-identifiable data from hospital and community/GP records will be linked and stored in a secure environment and used in research. A computerised process will be used to select the data so that identifiable information is not seen by anyone.

LTHT and TPP will agree a common 'key' (known as a 'salt', which is used to generate digests) via direct phone-call, using a random digit generator. The salt will only be used for this project. Next, an automated procedure will select cancer and control patient records at the Leeds Teaching Hospitals NHS Trust, LTHT. A digest for each record will be created using the agreed salt and OpenPseudonymiser (Section 2.1). Meanwhile, TPP will also run OpenPseudonymiser with the same salt to create digests for community / GP records that have been opted in to ResearchOne (R1) for use in such research projects.

LTHT and TPP will securely pass their digests to the University of Leeds Integrated Research Campus, IRC. Where the IRC can match the digests, this means that both R1 and LTHT have data for these records. TPP and LTHT will produce de-identified datasets from these records and securely deliver them to the IRC.

The IRC Data Services Team will use the digests to link the datasets from TPP and LTHT. They will run OpenPseudonymiser using a different salt to produce a new digest for each linked record. This means the digests on the linked data do not even relate directly to the digests that TPP and LTHT hold. The Data Services Team will also process the data to prepare it for research. They will store the data on secure IRC servers.

The IRC Data Services Team will grant access to the data for our research team. Access is dependent on signing an IRC User Agreement to abide by the ethical and legal requirements of the data. The data remains safely on IRC servers and is accessed and analysed remotely. This means that the data is isolated and a log is kept of who accesses it.

When the research team generate research outcomes and aggregated data that underpins these, they submit these to the IRC Data Services Team. The team check they do not contain identifiable information or in any other way break the ethical and legal requirements of the data. Approved outcomes and data then leave the IRC to inform the public through publications, presentations, public repositories, for example.

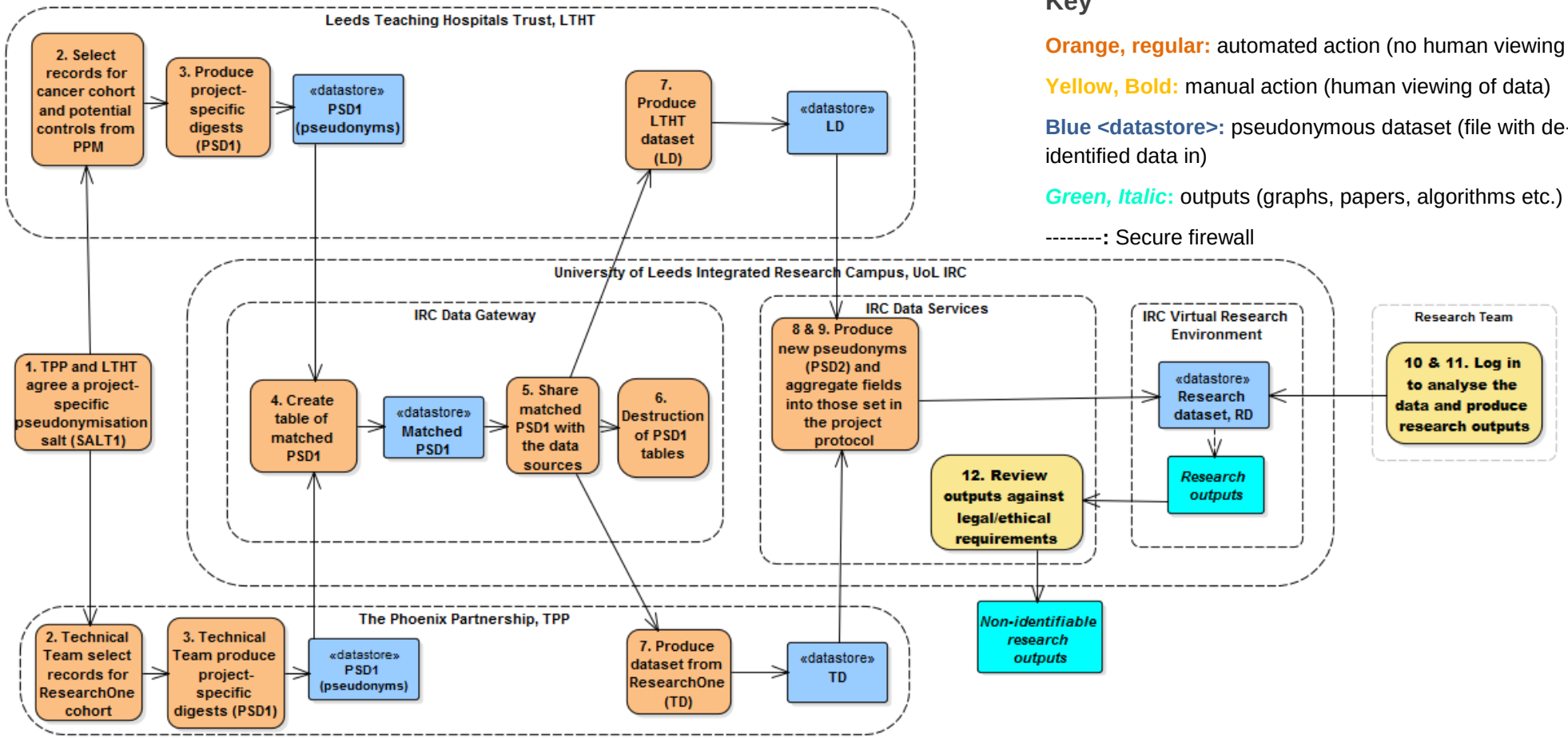
3. Technical Procedure

The following steps set out the data flow procedure and how OpenPseudonymisation will be utilised. **Figure 1** depicts these steps.

Further explanation is given in **Appendices 1 and 2** to illustrate the OpenPseudonymisation process. **Appendix 3** details where the data comes from. Please refer to **Appendix 4** for a glossary of abbreviations and definitions for OpenPseudonymiser terms such as “salt” and ‘hash’.

1. Leeds Teaching Hospitals NHS Trust (LTHT) and The Phoenix Partnership (TPP) agree a project specific “salt” phrase (using a random digit generator). They ‘hash’ this to create a hashed salt (SALT1).
2. LTHT run an automated procedure to identify Leeds patients with a cancer diagnosis or relevant non-cancer control.
3. LTHT use NHS number and date of birth (Year & Month e.g. 195010) as identifiers to generate a project specific digest (PSD1) for these records using OpenPseudonymiser and an agreed hashed project-specific salt (SALT1). This process is further explained in Appendices 1 and 2.
   TPP use NHS numbers and date of birth (Year & Month) as identifiers to generate a project specific digest (PSD1) for all ResearchOne records using OpenPseudonymiser and the agreed hashed project-specific salt (SALT1).
4. LTHT and TPP transfer the digests (PSD1) to the IRC Data Gateway, where the two digest lists are compared and a list of matching digests is compiled (Matched PSD1).
5. The matched digests (Matched PSD1) are provided to LTHT and TPP.
6. The digest lists (matched and original PSD1s) are deleted from the IRC Data Gateway.
7. LTHT produces a de-identified dataset (LD) for the project from the LTHT data warehouse in an encrypted file and shares it securely with University of Leeds Integrated Research Campus (UoL IRC) using “secure file transfer protocol”.
   TPP produces a de-identified dataset (TD) for the project from the ResearchOne database (R1). They will encrypt the file and share it securely with UoL IRC via secure upload (authenticated by IP address and login).
8. UoL IRC generate a second project specific digest (PSD2) using OpenPseudonymiser and a unique project-specific “salt” (SALT2) to replace PSD1 on LD and TD.
9. UoL IRC matches the datasets LD and TD, generates derived and aggregated data, placing the research dataset (RD) within a project specific research environment.
10. Named members of research team, approved by UoL IRC, are allowed "remote" access to the RD in the UoL IRC.
11. Research team generate research outputs.
12. UoL IRC Data Services screen and approve outputs against the ethical and governance requirements before they leave the UoL IRC.

Figure 1: Data flow for the project



Key

Orange, regular: automated action (no human viewing of data)

Yellow, Bold: manual action (human viewing of data)

Blue <datastore>: pseudonymous dataset (file with de-identified data in)

***Green, Italic*:** outputs (graphs, papers, algorithms etc.)

--------**:** Secure firewall

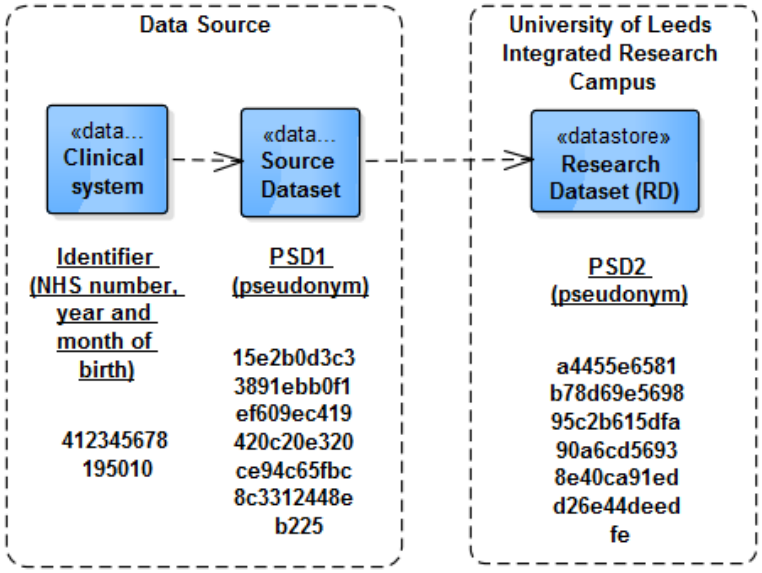
Summarised Glossary *Further terms are given in Appendix 4*

Salt: a random text phrase that is joined to NHS number and month/year of birth prior to OpenPseudonymisation

Digest: a 'pseudonym' created from the NHS number, month/year of birth and salt during OpenPseudonymisation

PPM and ResearchOne: electronic health record databases

Appendix 1. Worked Example of Pseudonymisation with a Record Identifier



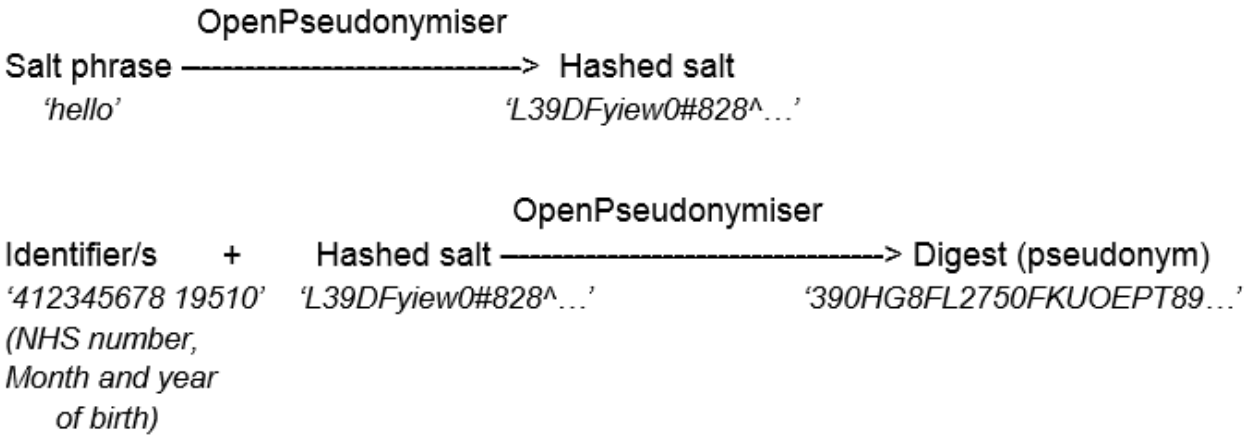
Key

--------: Firewall (approval is required to cross this boundary)

This diagram shows how a record identifier (in this case, the fictional NHS number '412345678' and year and month of birth '195010') would change through pseudonymisation. It shows three stages, which can be described as follows:

- The clinical systems at LTHT and TPP have electronic health records that hold the identifiers (NHS number).
- This is replaced by a digest (PSD1 *in Figure 1*) in the source-produced datasets (LD and TD *in Figure 1*).
- UoL IRC then replace PSD1 with a new digest (PSD2 *in Figure 1*) in the research dataset (RD *in Figure 1*).

Appendix 2. Worked Example of the OpenPseudonymisation Approach



This diagram shows how a project-specific digest (pseudonym) is produced, with a worked example. Please refer to Appendix 4 for a glossary of terms and definitions.

LTHT and TPP will use this process to produce PSD1 (*in Section 3*) from identifiers plus SALT1 (*in Section 3*), as follows:

- LTHT and TPP will download OpenPseudonymiser from the University of Nottingham.
- LTHT and TPP will agree on a project-specific "salt" phrase with which they will use OpenPseudonymiser to create a SHA-256 hashed "salt" (SALT1).
- This salt is effectively 'extra data' and the data sources use this plus an agreed identifier such as NHS number and OpenPseudonymiser to produce project-specific digests that act as pseudonyms (PSD1).
- LTHT and TPP will only use SALT1 to produce datasets for this project. They will not share SALT1 further without explicit ethical approval.

UoL IRC will use this process to produce PSD2 (*in Section 3*) from PSD1 plus SALT2 (*in Section 3*), as follows:

- UoL IRC will download OpenPseudonymiser from the University of Nottingham.
- UoL IRC will produce a project-specific "salt" phrase with which they will use OpenPseudonymiser to create a SHA-256 hashed "salt" (SALT2).
- LTHT and TPP provide UoL IRC with project datasets (LD and TD) that contain PSD1.
- UoL IRC will run SALT2 and PSD1 through OpenPseudonymiser to produce specific digests (PSD2) for the research dataset (RD)
- UoL IRC will only use SALT2 to produce research datasets for this project. They will not share SALT2 further without explicit ethical approval.

Appendix 3. Data Providers

The data required for this project is contained within hospital and primary care clinical and financial records. These longitudinal records include information about co-morbidity, immediate side-effects and late-effects of treatment, service utilisation, morbidity and mortality. Linkage of these two sources of data into a single, more comprehensive dataset will facilitate research.

A3.1 The Phoenix Partnership (TPP)

TPP (The Phoenix Partnership) is a major handler of health and social care records across the UK and conveniently co-located in Leeds. Over 5,000 health and social care organisations maintain health records using SystmOne, TPP's clinical system. These records are integrated through a 'cloud-like', approval-based access structure so that each patient (40 million patients in total) has a single record of care.

TPP maintain ResearchOne (R1), a database for research data provision with national ethics approval. There are 7 million non-identifiable patient records on R1. Health and social care organisations using SystmOne can opt-in to providing non-identifiable records data to R1 for use in approved research, and patients can opt out. There is a simple procedure for opting out patients from R1, which leads to removal from the database within one week. R1 provide posters and information leaflets to all participating organisations. Data is de-identified within SystmOne and transferred to R1 for use in projects approved by the programme of research. R1 was developed with patient and clinician engagement, deemed by the NIGB (now CAG) to hold de-identifiable data, and approved by the REC (11/NE/0184).

A3.2 Leeds Teaching Hospitals NHS Trust (LTHT)

The Leeds Cancer Centre is a key part of Leeds Teaching Hospitals Trust (LTHT) one of the largest teaching hospitals in Europe. There are around 2.5 million records on the main LTHT system, PPM+, including approximately 250,000 detailed records for patients with a cancer referral. LTHT have internally linked these clinical records to LTHT financial data.

Clinical teams within LTHT have used PPM and hospital financial data for clinical governance, local audit and more comprehensive research projects (with appropriate R&D and ethical approval). The LTHT Pseudonymisation Process is used to provide research datasets and is modelled on the ethically reviewed R1 de-identification procedure.

Appendix 4. Glossary of Terms and Definitions

Below is a **list of abbreviations** used in this protocol and their full terms:

Abbreviation	Full Term
LTHT	Leeds Teaching Hospitals NHS Trust
TPP	The Phoenix Partnership, provider of ResearchOne (and SystmOne electronic health records to primary, secondary and social care providers)
R1	ResearchOne (research organisation with opted-in records data)
UoL	University of Leeds
IRC	Integrated Research Campus (University of Leeds secure data services)
SALT1	Project specific "salt" used by LTHT and TPP to produce project specific digests
PSD1	Project specific digests produced by LTHT and TPP to enable pseudonymous linkage
LD	A project-specific de-identified dataset from LTHT
TD	A project-specific de-identified dataset from TPP
SFTP	"Secure file transfer protocol" is a network protocol for file transfer over a secure data connection
SALT2	Project specific "salt" used by UoL IRC to produce project specific digests
PSD2	Project specific digests produced by UoL IRC to provide the research team with patient-level digests in the research dataset (for the purpose of patient-level analysis) that do not link directly back to digests held by LTHT or TPP (PSD1)
RD	The research dataset viewed by the research team within the IRC Virtual Research Campus

Below is a **list of terms** and their definitions as used herein. These have been informed by the following documents:

- DD ISO/TS 25237:2008
- IRC Glossary of Terms
- LTHT Anonymisation / Pseudonymisation Process
- IRC Pseudonymous Linkage Operations
- HSCIC Code of Practice on Confidential Information
- ICO Anonymisation Code of Practice

Term	Definition
Identifiable data	Data that may reasonably be expected to include information that may identify a living or deceased individual, or may relate to an individual when combined with other fields either within the dataset or that may reasonably come into the possession of the approved data recipient.
Non-identifiable (or de-identified) data	Data in a field or dataset that does not enable an individual to be identified using reasonable effort. It may be created from

	identifiable data through the process of anonymisation or pseudonymisation.
Pseudonymisation	Processing that is applied to agreed fields within datasets to enable linkage between agreed datasets by means of a pseudonymous digest.
Digest	A digest is a 'pseudonym' created from agreed data fields during pseudonymisation.
Salt	A salt is a random text phrase that is joined to field(s) (such as NHS number) prior to pseudonymisation.
Hash	A cryptographic function used to change data in a way so that it is practically impossible to re-convert back.
SHA-256	Secure Hash Algorithm 256 is an industry-standard hash function.
Project specific salt	By adding a salt that is unique to each project to the fields being pseudonymised, the resultant digests differ for each project. This helps to prevent data being linked across project boundaries.
Project-specific digest	When a digest is created using a project-specific salt, it enables linkage between agreed datasets for an agreed purpose. Datasets with digests produced using a different salt would not link across.
Patient-level digest	This is a pseudonymous digest that is patient-specific. This enables a) the recipient to interpret the data at the patient level and b) the data provider to re-identify the patient for legal, health or safety purposes.
Anonymous data	Non-identifiable data that is not linkable to other datasets. Anonymous data should be handled in conditions that are appropriate to maintain its anonymity.
Pseudonymous dataset	A non-identifiable dataset that contains otherwise anonymous data linked to a pseudonymous digest. It can only be linked to datasets containing digests produced in the same way. Such data should be accessed and stored in conditions that are appropriate to maintain its non-identifiability.
OpenPseudonymiser	Pseudonymisation software used to produce project-specific encrypted salts and digests. It is a University of Nottingham application that utilises SHA-256 (Secure Hash Algorithm).