Data Linkage & Pseudonymisation Tool (DLPT)

Implementation Guide for Researchers and Data Managers

Authors:

Dr Polly Ashford, Research Lead (Service Delivery Interventions / Routine Data)

Martin Pond, Head of Data Management

This guide is designed to accompany the following DLPT templates:

NCTU DLPT Template_v1.0.xlsm

If you encounter any problems while following the instructions below, please contact dm.norwichctu@uea.ac.uk

Table of Contents

Introduction to the DLPT

The Norwich CTU DLPT template is a macro-enabled Excel workbook designed to process data exported from trial site patient management software.

The template can be adapted to the needs of specific clinical trials or other types of research using routinely collected data, where record linkage and pseudonymisation at site is required.

Once adapted, the DLPT can be used by site staff who have access to personal data in their normal duties (such as an NHS Trust data manager) to create an appropriately de-identified data file that can be transferred to the researchers. Data transfer may take place via cloud storage, email, or by direct import into a clinical trial database.

The DLPT has been designed to support trials using a REDCap trial database. Where possible, aspects of the tool which are specific to REDCap have been highlighted, and alternative approaches and workarounds suggested.

The original purpose of the tool was to allow clinical records to be linked with those of consented trial participants, while pseudonymising the data for those patients who have not consented to the trial, and whose data can therefore only be used in a de-identified format for research purposes.

If you have more specific requirements or additional complexity for the data linkage which is not covered by this guide, please contact Norwich CTU for advice: dm.norwichctu@uea.ac.uk

1 Adapting the DLPT Template

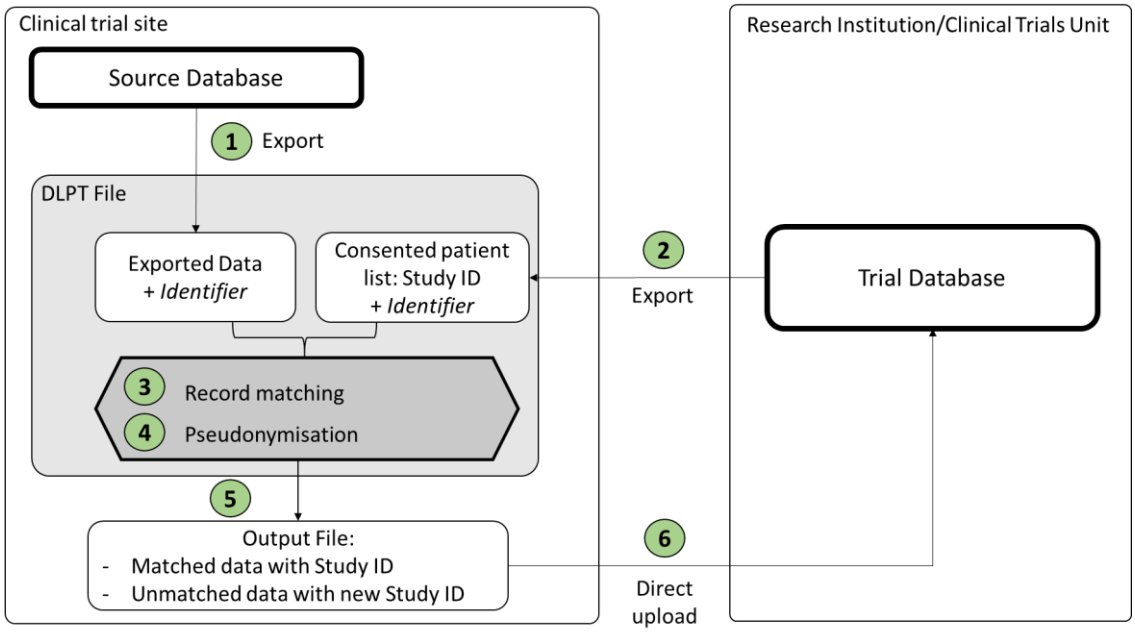
1.1 Introduction

The DLPT can be used as a data linkage tool to match records held by researchers with other source datasets, such as hospital or GP practice data. This is particularly useful if there are large numbers of records to match, which would otherwise be a time-consuming manual process.

In this scenario, the site staff download a list of participant IDs and link identifiers from the trial database, copy this into the DLPT along with the data output from their source database, and run the matching process.

A linked data file containing Study IDs but no identifiers can then be safely transferred back to the researchers via secure file transfer, uploaded to cloud storage or (where possible) directly imported into the trial database.

Unmatched records in the DLPT can be deleted prior to export, or if the study has ethical approval to use this data in a de-identified format, new study IDs can be applied by the DLPT to these records.



Example Scenario:

Trial Design: 2500 patients from 30 GP practices in England are being recruited to a randomised controlled trial to evaluate a staff training programme on asthma patient care. The primary outcome is the number of asthma annual reviews conducted over 12 months.

All patients are made aware of the trial using posters and social media, and a small sub-set (500 patients) are consented to complete baseline and follow-up questionnaires for secondary outcome measures.

For this trial, a large volume of routinely collected data will need to be obtained from patients' records, including asthma diagnosis, monitoring, and prescribing information, as well as non-asthma diagnoses, appointments and prescriptions.

Routine Data Collection and Linkage:

- A custom-built report in the GP practice clinical database is used to collate the data needed by the researchers. The results can be exported as a .csv file.
- Each row contains a strong identifier (NHS number) and also a patient software ID (which could be used to re-identify the data under some circumstances).
- The practice data manager will be given access to a report in the trial database listing the Study IDs and NHS numbers of all consented patients recruited at their site.
- The DLPT will be used to match consented patients with patients in the GP data, using NHS number as the link identifier, and replacing the patient software ID with the study ID.
- Unmatched patients will be assigned a new Study ID in the format required.
- The DLPT will then remove NHS number and export the data to a de-identified .csv output file.
- Finally, the practice data manager will directly upload the output file into the trial database.

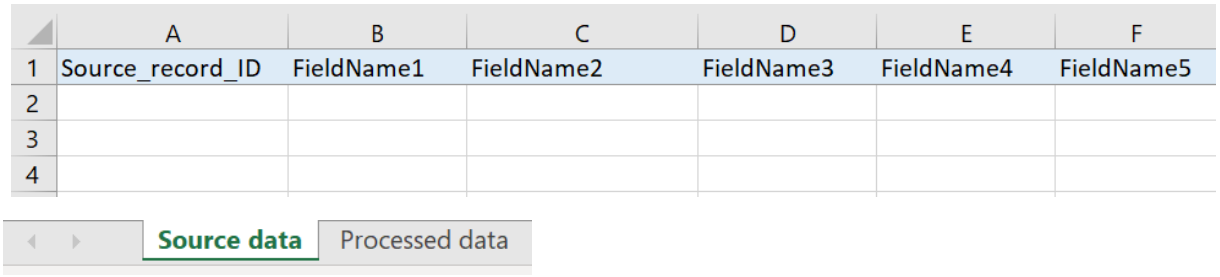
1.2 Template set-up instructions

Follow the steps in Section 1.2.1 to 1.2.6 to set up the DLPT for use at a site. Cells highlighted in pale blue indicate cells to be edited during the set-up process.

1.2.1 Sheet 1: Source Data

'Source data' is the sheet onto which the site user will paste the source data.

1. To prevent errors, pre-populate row 1 with the field names of the source dataset specification. The template DLPT contains placeholder fields as follows:

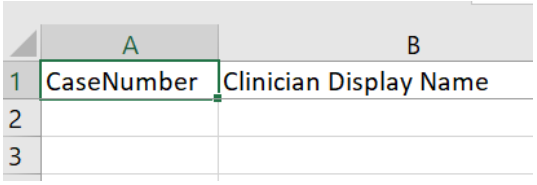


	A	B	C	D	E	F
1	Source_record_ID	FieldName1	FieldName2	FieldName3	FieldName4	FieldName5
2						
3						
4						

Source data | Processed data

Note that column A (and therefore the first column in the source data export) must be the record ID of the source database. This should not be the link identifier used to match participant records.

e.g.



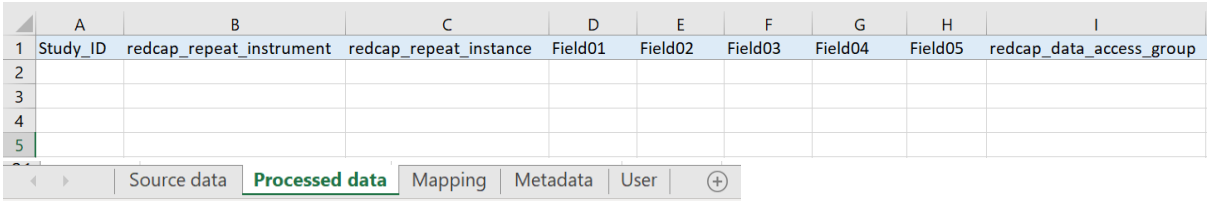
	A	B
1	CaseNumber	Clinician Display Name
2		
3		

2. Protect the first row of this sheet to prevent it being overwritten.

1.2.2 Sheet 2: Processed Data

'Processed data' is the sheet that will contain the processed data ready for export when the data linkage and pseudonymisation has taken place.

1. Pre-populate row 1 with the field names required for export from site to researcher.
   e.g.



	A	B	C	D	E	F	G	H	I
1	Study_ID	redcap_repeat_instrument	redcap_repeat_instance	Field01	Field02	Field03	Field04	Field05	redcap_data_access_group
2									
3									
4									
5									

Source data | Processed data | Mapping | Metadata | User

   Note that the following columns are fixed:
   - Column A will always be populated with the new study ID, replacing the source data record ID, and should be labelled as such.
   - Columns B and C must be labelled with the fields required for a repeat instrument in a REDCap trial database, as shown. The last column on this sheet must **If you are not using**

REDCap, or your REDCap database does not contain repeat instruments, contact dm.norwichctu@uea.ac.uk for a workaround.

- The last column must be labelled 'redcap_data_access_group' as shown. This refers to the site identifier – more information on this is given in Section 1.2.5.

Depending on the trial, the field names between column C and the final column may simply match the source data. However, the following should be considered:

- **Is the processed data intended to be imported into a REDCap trial database?**
  If so, the field names must match existing fields in the trial database for the import to be successful. The correct format can be obtained by downloading the template import file from the REDCap project (see Appendix 1).
- **Alternatively, if the data is being collated manually by the researchers**, and not directly imported into a database, it may still be more efficient to convert the field names via the DLPT to standardise them. For example, if the data is being obtained from sites using different software systems.

2. Protect the first row of this sheet to prevent it being overwritten.

1.2.3 Sheet 3: User

'User' is the sheet where the site user will enter the Study IDs and link identifiers of consented participants from the trial database. See Appendix 2 on how these can be accessed by the site.

1. Cells A1 and B1 should be pre-populated with the relevant field names:
   - Column A must be the identifier field (e.g. NHS number)
   - Column B must be the existing Study ID.

	A	B
1	Identifier	Study_ID
2		
3		
4		

E.g.

	A	B	C	D	E	F	G
1	NHS_number	record_id				Run	Reset
2							
3							

Note: Macro buttons 'Run' and 'Reset' in cells F1 and G1 are used by the site to run the data linkage process and should not be edited.

1.2.4 Sheet 4: Mapping (Hidden)

'Mapping' is the sheet that is used to map the field names on 'Source Data' with the corresponding fields on 'Processed data'.

1. In cell A2, select the field name of the link identifier from the dropdown menu, and in cell B2, select the study ID field:
   e.g.

	A	B
1	Columns in Source data	Columns in Processed data
2	NHS number	Study_ID
3		

2. From cell A3 down, select each field name from the dropdown menu, and select the corresponding field names in column B.

	A	B
1	Columns in Source data	Columns in Processed data
2	NHS number	Study_ID
3	FieldName2	Field02
4	FieldName3	Field03
5	FieldName4	Field04
6	FieldName5	Field05

Note that the template has been set up to process a maximum of 266 columns/fields. If you require more fields, extend the formulae in columns C and D ('IndexFrom' and 'IndexTo') as needed.

3. In cell F2, select the link identifier field from 'Source data', such as NHS number, e.g.

F	G	H
Identifier in Source data	Sequence field in Source data	Target instrument
NHS number	FieldName3	Routine_Data

4. In cell G2, select the field from 'Source data' containing a sequential value that can be used to sort the data:
   - If the data contains multiple rows per participant, if may be appropriate to use appointment date to sort the records.
   - If the data contains one row per participant, the data can be sorted on any sequential value, such as year of birth, date of registration etc.

   Note: a value must be selected for the DLPT to proceed.

F	G	H
Identifier in Source data	Sequence field in Source data	Target instrument
NHS number	FieldName3	Routine_Data

5. In cell H2, enter the name of the REDCap 'target instrument' where the data will be uploaded to.

F	G	H
Identifier in Source data	Sequence field in Source data	Target instrument
NHS number	FieldName3	Routine_Data

1.2.5 Sheet 6: Metadata (Hidden)

'Metadata' contains information needed to apply the new study IDs to records, according to the desired format.

The DLPT was created for multi-site studies, and assumes that a site identifier forms part of the study ID. E.g. **SP01-001** where **SP01** = site and **001** = participant.

1. In cell I2, enter the site identifier as it appears in the Study ID. If using a REDCap database, this is likely to be the Data Access Group (DAG) name.

e.g.

H	I	J	K	L
Seed	redcap_data_access_group	__GROUPID__	Max	Next
500	Site_01		0	501

2. In cell H2, enter the value above which sequential numbering of unmatched Study IDs should start.

H	I	J	K	L
Seed	redcap_data_access_group	__GROUPID__	Max	Next
500	Site_01		0	501

In the example shown, the 'Seed' value is set to 500. If the maximum existing study ID is equal to or lower than 500 on 'User', the next unmatched record will be assigned 501. If the maximum existing study ID is above 500, the DLPT will add 1 to the highest value to prevent duplicate values being assigned.

1.2.6 Finalising the DLPT

Before sending the DLPT to site, it is recommended that you do the following:

- Protect and hide sheets 'Mapping' and 'Metadata'
- Save the DLPT file with a site identifier in the file name as appropriate
- Prepare user instructions for the site, using the template provided

1.2.7 Unmatched records

The DLPT is designed to apply new Study IDs to unmatched patient records, where the study has ethical approval to collect de-identified data for all patients included in the source data export. If this is not the case, the site user should be instructed to filter and delete the unmatched records prior to sending data to the research team.

2 Data Transfer from site to researcher

2.1 File sharing

The DLPT creates a .csv file of the data on sheet 'Processed data' in the final step.

It is recommended that the following the instructions above, and with consideration of the data export specification, the DLPT is set up such that this file does not contain identifiable information.

The file can then be safely uploaded to a securely shared folder (e.g. OneDrive) for further processing by the researchers.

2.2 Direct import into a trial database

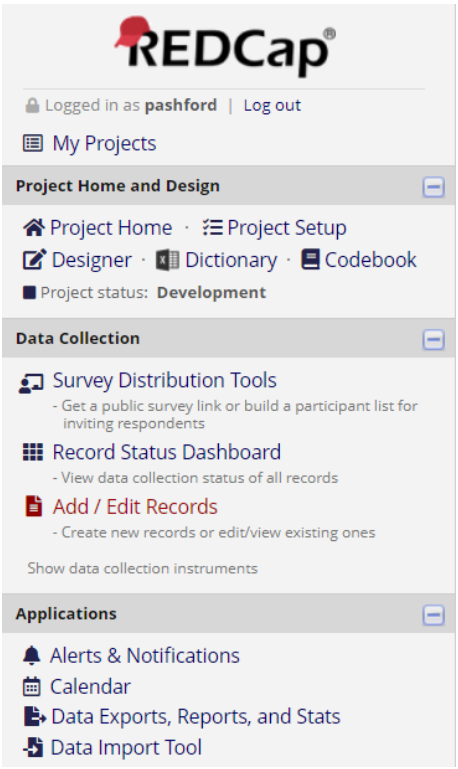
The DLPT was designed with a REDCap clinical trial database in mind. If the file has been set up in line with the instructions above, it is possible to grant the site user permissions to import the exported .csv file directly into the REDCap trial database. The researchers should consider the implications of this for the data management of their trial, e.g. potential for data to be overwritten, pre-import checking requirements etc.

Appendix 1: REDCap data import specification

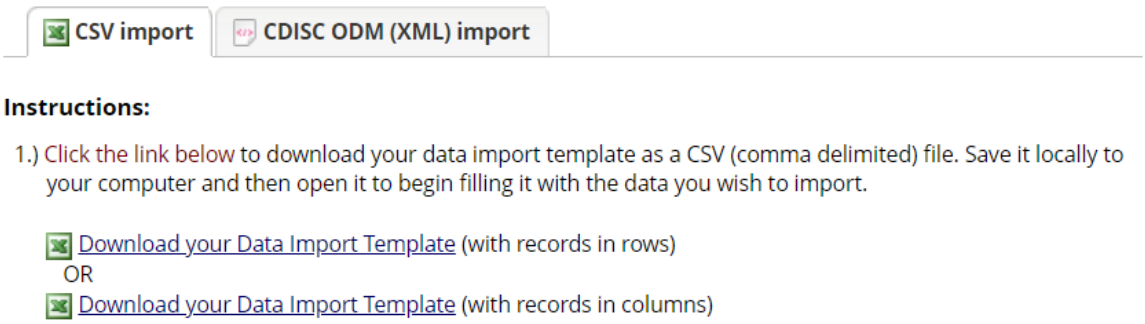
If your processed data .csv file is intended to be imported directly into a REDCap trial database, the field names in the 'Processed data' tab of the DLPT file must match fields in the REDCap database.

To ensure smooth import of data into REDCap, check the import specification by downloading the Data Import Template from the REDCap project as follows:

1. Go to 'Data Import Tool':



2. Click 'Download your Data Import Template (with records in rows)':



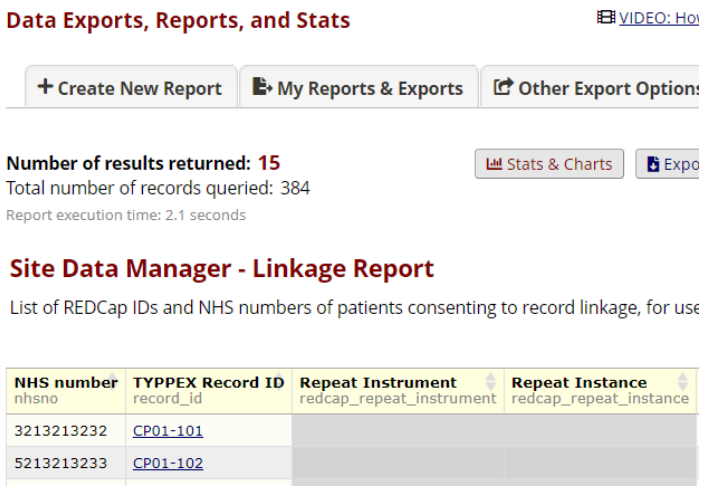
3. Use these field names to pre-populate the 'Processed data' sheet in the DLPT. **Note that any fields from the database which are unrelated to the site source data, such as online participant surveys/e-consent etc, should not be included in the DLPT.**

Appendix 2: Site user access to trial database

Option 1: REDCap trial database

If the trial database is held in REDCap, the site user can be given an account and access to the limited amount of information needed to run the DLPT. A site-specific report can be used to create the consented participant list needed to populate the 'User' tab of the DLPT.

E.g.



The site user can be instructed to select, copy and paste the required fields (i.e. column 1 and 2) or export the data to CSV before copy and pasting the data, if there are large numbers of records.

Option 2: Alternative approach

It may be that a different trial database software is used, and/or it is not possible to give the site user direct access to the required information in the database.

The researcher can alternatively make the study IDs and link identifiers of consented participants available to the site user via a securely shared file such as an Excel document held in an access-controlled OneDrive folder.