



UK Longitudinal Linkage Collaboration

Population Health Sciences

Bristol Medical School

Canynges Hall

39 Whatley Road

Bristol BS8 2PS

UK Longitudinal Linkage Collaboration (UK LLC)

Trusted Research Environment (TRE) User Guide

PUBLIC

Version 2.1

26 February 2024

This document was printed on: Mon, 26 Feb 2024 and may be out of date. The current version can be found at: [Data Team](#)

Page 1 of 26

CONTENTS – UK LLC TRE USER GUIDE

1. INTRODUCTION	3
1.1 Video help	3
2. UK LLC DATA FLOWS	4
3. SETTING UP AND LOGGING IN	5
3.1 Account setup and first-time login	5
3.2 Direct log-in	7
4. WORKING IN THE TRE	7
4.1 Data introduction.....	7
4.2 Project folders and project naming	7
4.3 Retrieving data – ODBC connection	8
4.4 Metadata	9
4.5 GitLab/GitHub.....	9
4.6 Data refreshes	14
4.7 Reporting concerns about data	14
4.8 Detecting and preventing scope creep.....	14
4.9 Audits	15
5. USING SOFTWARE IN THE TRE	15
5.1 Software available	15
5.2 Using SQL Server Management Studio	15
5.3 Using Python.....	16
5.4 Using R	17
5.5 Using Stata	18
6. CORE DATA FILES	20
6.1 CORE denominator file1	20
6.2 CORE NHSD Presence	20
6.3 CORE NHSD Derived indicator	21
6.4 CORE NHSD LSOA11.....	21
6.5 CORE LSOA11 geo indicators	21
6.6 NHSD DEMOGRAPHIC SUB	22
7. WELLCOME TRUST COVID QUESTIONNAIRE LOOKUP.....	22
8. GETTING FILES INTO THE TRE.....	23
9. GETTING FILES OUT OF THE TRE.....	23
9.1 Introduction	23
9.2 File-out process	23
9.3 File-out rules.....	24
10. AMENDMENTS TO PROJECTS.....	24
11. FINISHING A PROJECT	25
11.1 Request final file outs	25
11.2 Share your syntax.....	25
11.3 Share your derived data outputs.....	25
12. PUBLISHING OR PRESENTING YOUR WORK	25
13. USER SUPPORT	26
13.1 UK LLC Data Users' Google Group	26
14. MAKING A COMPLAINT.....	26

1. Introduction

[UK Longitudinal Linkage Collaboration \(UK LLC\)](#) integrates data from UK Longitudinal Population Studies (LPS), linked to health, environmental and administrative records, into a centralised research resource available in a Trusted Research Environment (TRE). The UK LLC TRE is hosted by Swansea University's [Secure eResearch Platform \(SeRP UK\)](#).

This User Guide provides guidance for accessing and working in the UK LLC TRE and is available to all prospective and current UK LLC researchers. This guide explains how to:

- Set up and log into the UK LLC TRE
- Access and work with your approved data, including the various software packages available
- Use the UK LLC Git repositories
- Interpret standard UK LLC datasets
- Bring files in and out of the TRE, including data output guidance
- Request an amendment to your project
- Close down your project
- Publish or present your analyses.

This guide should be read in conjunction with the [UK LLC Data Access and Acceptable Use Policy](#) which details the terms and conditions under which researchers agree to access and use data held in the TRE. Researchers should also refer to the [UK LLC TRE FAQs](#) and UK LLC's suite of policies, including:

- [UK LLC Reproducible and Reusable Research Policy](#)
- [UK LLC Naming of Projects and Data-related Outputs Policy](#)
- [UK LLC Output Review Policy](#)
- [UK LLC Project Amendment Policy](#)
- [UK LLC Publication Policy](#).

To access data in the UK LLC TRE, researchers must have current ONS Approved Researcher Accreditation, have an approved project and have completed all necessary UK LLC paperwork (Data Access Agreement and Data User Responsibility Agreement).

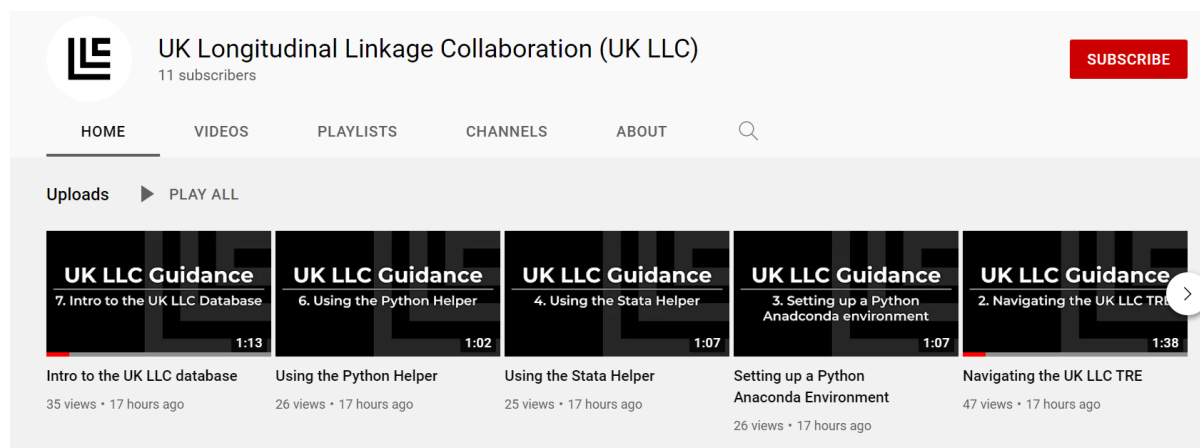
NOTE: This document is under development. Therefore, some sections are not complete and are marked as such.

1.1 Video help

Much of this document has accompanying videos. The videos cover:

- An introduction to the UK LLC TRE, including logging in and navigating the TRE
- Setting up a Python environment
- Using helper tools.

The videos are available via the [UK LLC YouTube channel](#). By clicking on "PLAYLISTS" and "PLAY ALL" the videos will play in sequential order.

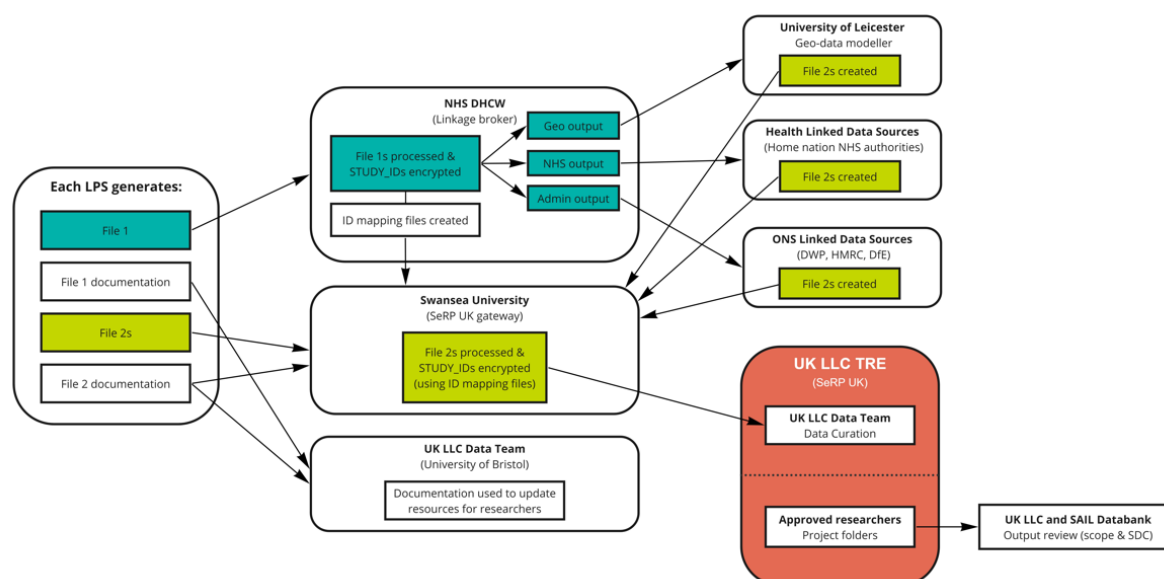


2. UK LLC Data Flows

See the data flow diagram below, which gives a high-level overview of the data processing methodology used to flow data into the UK LLC TRE.

Data linkage and extraction are managed through a Trusted Third Party (NHS Digital Health and Care Wales, DHCW). Flows of data from contributing LPS and the owners of routine records are conducted through a 'split file' protocol where the flow of identifiers is entirely separate from the flow of attribute data. Data integration and management is conducted by a dedicated UK LLC Data Team within the TRE.

File 1s, containing only identifiers, are sent by LPS to DHCW where they are processed and STUDY_IDs are encrypted. An ID mapping file containing the original STUDY_ID (provided by LPS) and the encrypted STUDY_ID (created by NHS DHCW) is sent to Swansea University (SU - this enables SU to bring together the various attribute data for each participant). DHCW act as the linkage 'broker', facilitating linkages by sending a reformatted and permission-filtered file of unique identifiers and encrypted STUDY_IDs to NHS England and University of Leicester (UoL) for processing. NHS England and UoL then create and send 'File 2s' containing de-identified attribute data, including the encrypted STUDY ID, to SU for encryption of the STUDY_IDs and ingest into the UK LLC TRE. File 2s (created by LPS) are sent to SU where, as for the other File 2s, STUDY_ID is encrypted during ingest to the UK LLC TRE.



NOTE: There are ongoing discussions to flow Scottish (Public Health Scotland), Welsh (NHS DHCW/SAIL) and Northern Irish (Health & Social Care Northern Ireland) NHS records into the UK LLC TRE. The UK LLC is also seeking agreement to link to administrative records from the Department for Work and Pensions (DWP), HM Revenue and Customs (HMRC) and Department for Education (DfE) via the Office for National Statistics (ONS).

3. Setting Up and Logging In

3.1 Account setup and first-time login

The UK LLC TRE is based on the SeRP UK infrastructure. In order to setup and log into your account, you will need the following (provided by the UK LLC Data Team upon approval of your application and completion of paperwork):

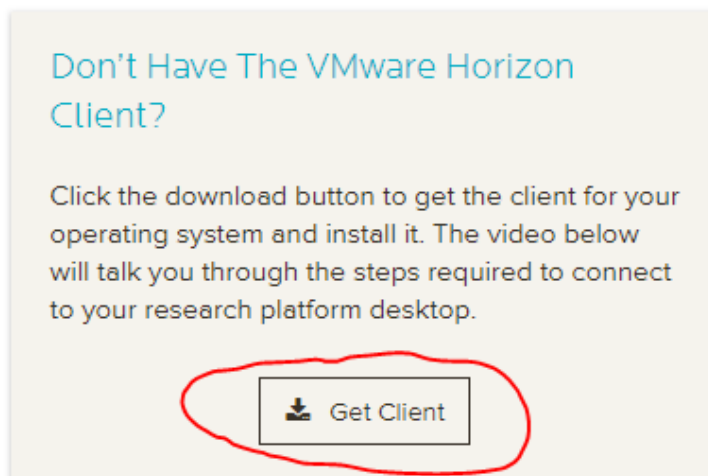
- Username
- QR code for 2-factor authentication (2FA).

When you have received your username and QR code, follow the steps below to **set up your account**:

1. Navigate to: <https://portal.ukllc.ukserp.ac.uk/requestnewpassword> and enter either your username or email address.
2. You will then receive an email with a link to set up your password. Follow the link and enter your password (twice). Please do not share your password with anyone else.
3. QR code: open/download a “One-Time Password” (OTP) app on your smartphone (recommend FreeOTP) and scan the QR code provided by the UK LLC Data Team. Please keep the QR code secure and do not share it with anyone else. Next type the unique 6 digit code generated by the OTP app in the 2FA passcode field. Finally submit the form.

When your account is set up, follow the steps below to **log in**:

1. Navigate to: <https://portal.ukllc.ukserp.ac.uk/> and enter your username and password (use Google Chrome). At your first login, you will need to download VMware Horizon Client from the right side of the webpage:



2. Once installed, you will be able to launch the SeRP, click ‘Launch Platform’:



Connect to the Research Platform


To connect to the research platform you must have the VMware Horizon Client installed on your machine.

Remember to have your YubiKey/Mobile Authenticator ready.



Click 'Open VMware Horizon Client' from the pop-up box and 'Accept' the terms when prompted.

3. The following dialogue box appears:

 <https://connect.serp.ac.uk>


Enter your 2FA user name and passcode.

Passcode

CancelLogin

Enter username and with cursor in 'Passcode' field generate an OTP with your app and enter the code into the form.

4. The following dialogue box appears:

 <https://connect.serp.ac.uk>

Enter your password


CancelLogin

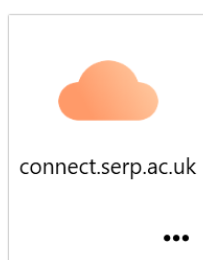
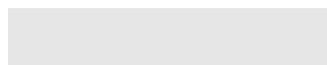
Enter your password, and the desktop should launch.

When you have finished in the UK LLC TRE, save your work, select 'Options' and then 'Log off' from the top bar menu.

3.2 Direct log-in

Once your account is setup, you can also access the UK LLC TRE directly via VMware, rather than opening the portal and launching from there. To do this, navigate to My Computer and launch VMware Horizon Client. Click 'connect.ukserp.ac.uk' and follow steps 3 and 4 from section 3.1 above.

 VMware Horizon Client



4. Working in the TRE

4.1 Data introduction

Data are stored on a SQL Server relational database which sits within the UK LLC TRE. The database name is 'UKSERPUKLLC'. The UK LLC data provision pipeline is designed to provision a bespoke set of 'views' of the 'tables' held in the database to researchers, based on the data requested for their specific project. A database view provides a tailored, read-only, live representation of the underlying data (stored in the database tables). Using views enables the UK LLC to customise your data provision, conduct internal linkages and implement governance controls. Within relational databases, groups of tables and views are organised into 'Schema'.

Each 'view' of data provided is named following the convention: 'LLC_XXXX.SCHEMA_name_vXXXX_yyyyddmm', where:

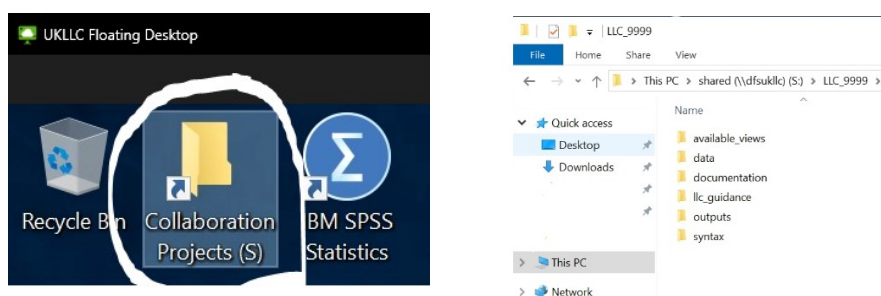
- 'LLC_XXXX' is the project number
- 'SCHEMA' is the provenance (e.g. the name of a specific LPS, linked data source)
- 'name' is the block name (as selected on the data request form)
- 'vXXXX_yyyyddmm' is the version number and date associated with the version.

Associated metadata (value and variable labels) are also provided as database views. A 'Codelist Library' of all codelists requested for any project to date are available as 'tables' for researchers' reference.

4.2 Project folders and project naming

Each UK LLC project is allocated a 'project folder' within the UK LLC TRE. The project folder is given the same unique project number that was assigned by the Applications Team when the application to access the TRE was submitted, e.g. llc_9999. It is important for public/participant transparency, the reusability of content, and governance compliance that these project numbers are used in a clear and consistent way across the project lifecycle. **Consequently, all research users must follow the [UK LLC Naming of Projects & Data-related Outputs Policy](#).**

In order to access your project folder, open the 'S drive':



Find your project folder, named with your UK LLC project number (e.g. llc_9999).

Your project folder is a secure working area, which can only be accessed by you and other approved researchers on your project and authorised UK LLC and SeRP UK staff. Your project folder should be used to store all workings related to your project (i.e. syntax, documentation, data files).

Project folders have a defined standard folder structure (as detailed below) to help guide the organisation of projects. We ask that these folders are used in a systematic manner to aid data curation and compliance. This is an important part of UK LLC's ability to maximise the reuse of research outputs in future research. However, please feel free to create additional sub-folders in your project folder as necessary for your research.

Sub Folder	Purpose
available_views	Contains csv file outputs from all data provisions with a list of SQL database views that have been made available to the project
data	For storing researcher generated datafiles created during analyses
documentation	For storing documentation pertinent to the research. This can be generated within the TRE or sent in via 'file-in' request
llc_guidance	A sub-folder containing key UK LLC requirements documents for ease of reference (e.g. your Data Request Form)
outputs	For storing proposed publication-ready analytical outputs to be submitted through the 'file-out' review process
syntax	For storing researcher generated analytical syntax/scripts

4.3 Retrieving data – ODBC connection

IMPORTANT: Your project data are not provided as files; instead they are available as database views. Therefore, you will not find data files in your project folder on initial login. Data will instead need to be retrieved from the **UKSERPUKLLC** database using an ODBC connection (a piece of software which enables programmes – such as R or STATA – to connect to databases and interact with the data).

There is a system ODBC Data Source available to all users in the TRE. This will allow connection to the UKSERPUKLLC database on which all data are held.

The helper scripts (python, R, Stata) use this data source to pull the relevant data for your project. If you prefer to query the data via another method or software package, the data source is called **'LLC_DB'**.

A list of your database views can be found in the 'available_views' sub-folder of your project folder.

4.4 Metadata

All metadata are provided as database views in schemas accessible to everyone in the TRE. Metadata are split into value labels called '**VALUE.all_values**' and variable labels called '**DESCS.all_descriptions**'. Using the variables 'table_name' and 'TABLE_SCHEMA', metadata can be linked to your data. Like all data, metadata can be viewed, queried and linked to your data using different software packages as described in section 5.

Value and variable labels are available for LPS data and the majority of NHS England data. Full labelling will be completed as part of an upcoming work package. In the interim, please use: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services>.

4.5 GitLab/GitHub

4.5.1 Introduction

Researchers must make their project workings (syntax, code lists, protocols, methods and derived variables) accessible and understandable to future users.

Each project is allocated a GitLab subgroup within the UK LLC TRE. Users should add cleaned and documented content to their own GitLab project space during the course of their project. The GitLab project/repository is named as your project number e.g. LLC_0000. This space is only accessible to analysts named on your project and to UK LLC staff.

There is another project/repository where all researchers have read (pull) access. This is called "ReadAll" and is a space for sharing code between projects. Only UK LLC data managers can write (push) to this project. Additional to this, each project has its own [GitHub](#) (external to TRE) repository named as their project number, e.g. LLC_0000. This is created during project setup with write (push) permissions included for all project members. **The internal GitLab "ReadAll" project should mirror the contents of the external GitHub.**

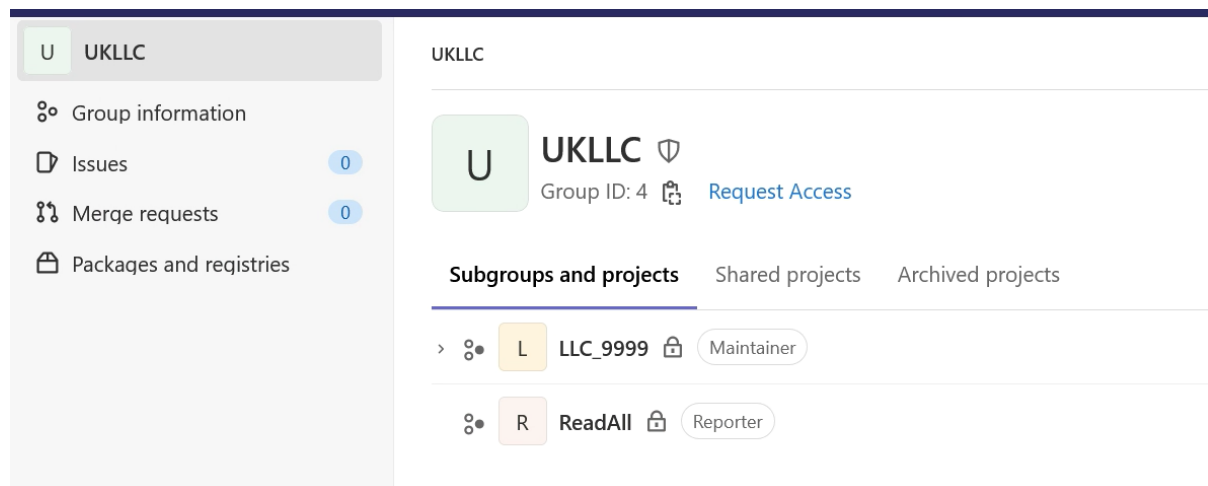
All research users are required to follow the [UK LLC Naming of Projects and Data-related Outputs Policy](#) for labelling research products.

If you are new to Git/GitLab/GitHub, we recommend reading the [GitLab documentation](#), especially the section covering [Git on the command line](#). There are also a number of good tutorials on Git and GitLab on YouTube. You'll find GitLab available on all UK LLC SeRP desktops. Here your permissions will already be mapped to the appropriate group/subgroups.

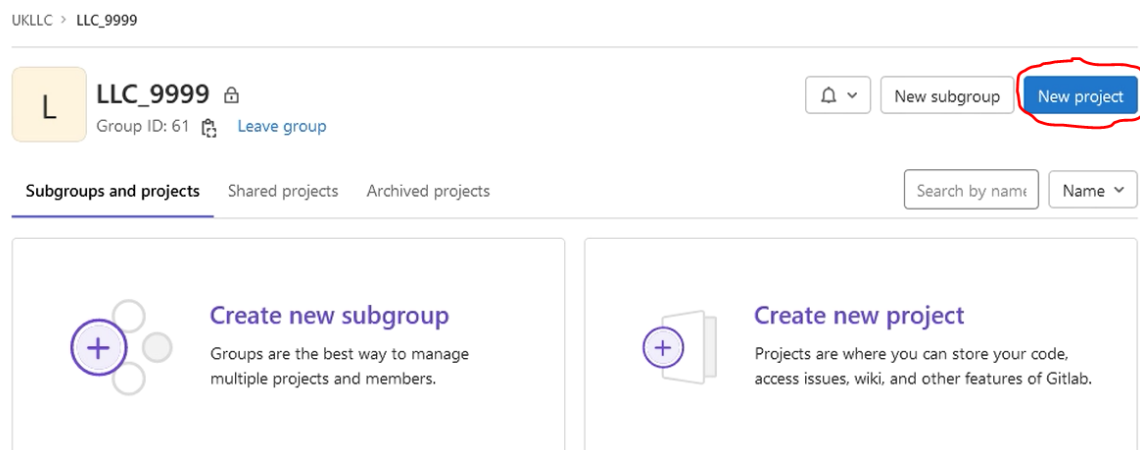
4.5.2 Accessing GitLab and creating a repository

To access GitLab within the SeRP open a browser (recommend Mozilla Firefox) navigate to <https://gitlab.ukllc.ukserp.ac.uk> or click on the "UKLLC GitLab" icon from the SeRP desktop.

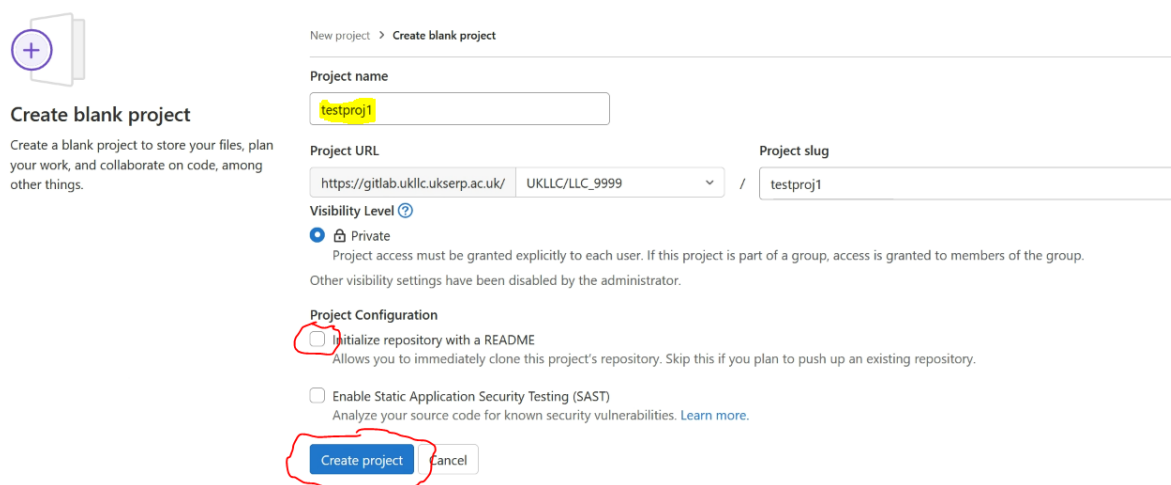
1. From main menu select "Groups" > "View all groups", then select "UKLLC".
2. Here you should see the Subgroups you belong to. This should include all projects you belong to and "ReadAll":



3. To create a version-controlled area within your project space you'll need to create a "project" (known as a repository in Git) within your subgroup.
4. To do this, click on the subproject you want to version control within (in this case LLC_9999) and click "New project":



5. Select "Create blank project".
6. Enter project name, this should follow the naming convention <project_number>_<descriptive_name> (e.g. "LLC_9999_GP_asthma_diag"), deselect "initialize repository with a README" and click "Create project":



- If this is your first time using GitLab, you'll need to set up an SSH key pair. See [guide](#) on how to do this. Having an SSH key set up is essential in allowing you to push to and pull from repositories.
- Once SSH keys are set up, decide/create folder in your project working space (Windows Explorer) containing files for version control. This folder should contain syntax and documentation only, not data. Data should be stored in a separate folder.
- Open Git BASH and cd (change directory) to folder "`S:\#insertProjectName#\#insertFolderName#`", e.g.:

```
tesri@uk11c-float-18 MINGW64 ~
$ cd "S:\LLC_9999\t1"
```

- To initialise Git within this folder type git init:

```
tesri@uk11c-float-18 MINGW64 /s/LLC_9999/t1
$ git init
```

- You'll need to add a "remote" to tell Git which remote repository in GitLab is tied to the specific local folder on your computer. The remote tells Git where to push or pull from. For this you'll need a project path.
- To get your project path go to the project page in GitLab, click the 'Clone' drop down and copy the address from 'Clone with SSH' cell:

UKLLC > LLC_9999 > testproj1

The screenshot shows the GitLab interface for a project named 'testproj1'. At the top, there are buttons for 'Find file', 'Web IDE', and a 'Clone' dropdown menu. The 'Clone' menu is open, showing three options: 'Clone with SSH', 'Clone with HTTPS', and 'Open in your IDE'. The 'Clone with SSH' option is highlighted in yellow, showing the URL 'git@gitlab.uk11c.ukserp.ac.uk:UKL'. Below this, the 'Clone with HTTPS' option shows the URL 'https://gitlab.uk11c.ukserp.ac.uk'. The 'Open in your IDE' section lists 'Visual Studio Code (SSH)', 'Visual Studio Code (HTTPS)', 'IntelliJ IDEA (SSH)', and 'IntelliJ IDEA (HTTPS)'. The 'Clone' button is circled in red.

Once you have this, in Git BASH type: git remote add origin `#insertProjectPath#`:

```
tesri@uk11c-float-18 MINGW64 /s/LLC_9999/t1 (master)
$ git remote add origin git@gitlab.uk11c.ukserp.ac.uk:UKLLC/LLC_9999/testproj3.git
```

- Switch branch to main with:

```
tesri@uk11c-float-18 MINGW64 /s/LLC_9999/t1 (master)
$ git branch -M main
```

- To add all files in the folder use: "git add ." or to add specific files use: git add `#specificFilename#`, e.g.:

```
tesri@uk11c-float-18 MINGW64 /s/LLC_9999/t1 (main)
$ git add testfile1.txt
```

- To commit changes: git commit -m "`commitMessage`":

```
tesri@ukllc-float-18 MINGW64 /s/LLC_9999/t1 (main)
$ git commit -m "first commit"
```

16. To push your local commits to the main branch of the origin remote:

```
tesri@ukllc-float-18 MINGW64 /s/LLC_9999/t1 (main)
$ git push -u origin main
```

4.5.3 Dos and don'ts

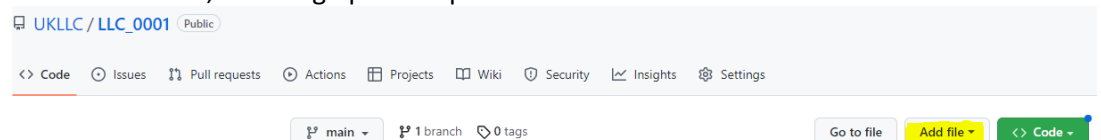
- Files in the repository should follow the [UK LLC Naming of Projects and Data-related Outputs Policy](#)
- Repositories themselves should follow the naming convention: <project_number>_<descriptive_name> (e.g. "LLC_9999_GP_asthma_diag")
- Please keep data files outside your repository and store in a separate folder. This is because these files cannot be shared between projects.

4.5.4 Adding files to ReadAll GitLab and external GitHub

As explained in the introduction, to share non-disclosive outputs, e.g. code and documentation with researchers from other projects in the TRE and to add to the public external GitHub, researchers must:


1. Request file-out and tick the box marked 'GitHub/GitLab sharing'.
2. The files will be disclosure checked.
3. Once cleared, UK LLC staff will copy the contents to the GitLab (internal to the TRE).
4. The **researcher must** now push the contents of the file-out to their project repository on the **UK LLC GitHub** and make them public (initially your repository is set to Private – see point 5 below). This can be done in a number of ways but the most simple is a drag-and-drop/file upload. To do this:

- a. Navigate to <https://github.com/UKLLC/> and sign in
- b. Find your project repository e.g. LLC_0000 and click to open
- c. Select "Add file", selecting option "Upload files":




- d. Drag/choose files, adding a commit message where highlighted before selecting "Commit changes":

LLC_0001 /



Drag files here to add them to your repo
Or [choose your files](#)



Commit changes

Add files via upload

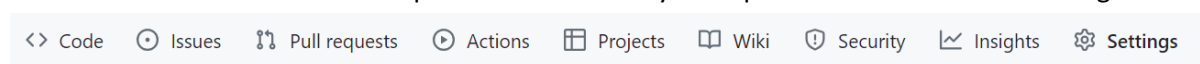
Add an optional extended description...

☒ Commit directly to the `main` branch.
☐ Create a new branch for this commit and start a pull request. [Learn more about pull requests.](#)

Note: your project repository will have a pre-populated readme file with important project information e.g. lay summary. Please feel free to add to this readme to give context to your files.

5. To set your **repository to Public** visibility:

- Navigate to the main page of the repository.
- Under your repository name, click Settings. If you cannot see the "Settings" tab, select the dropdown menu above your repo name and then click Settings.



- In the "Danger Zone" section, to the right of "Change repository visibility", click Change visibility.
- Select Public visibility.
- To verify that you're changing the correct repository's visibility, type the name of the repository you want to change the visibility of.
- Click I understand, change repository visibility.

4.5.5 Cloning a repository to your working area

If you wish to use code developed by another researcher you need to clone the repository from the ReadAll GitLab:

- Within the SeRP navigate to <https://gitlab.ukllc.ukserp.ac.uk> or click on UKLLC GitLab from the SeRP desktop. Recommended browser: Mozilla Firefox.
- From main menu select "Groups" > "View all groups", select "UKLLC", then select "ReadAll".
- Here you will see all repositories returned to UK LLC, select the repository you want to clone and click on the Clone dropdown and copy the text under "Clone with SSH":

UKLLC > ReadAll > testproj6

The screenshot shows the GitLab web interface for a project named 'testproj6'. At the top, there are buttons for 'Find file', 'Web IDE', and 'Clone'. The 'Clone' button is circled in red. A dropdown menu is open from the 'Clone' button, showing three main sections: 'Clone with SSH', 'Clone with HTTPS', and 'Open in your IDE'. Under 'Clone with SSH', the URL 'git@gitlab.ukllc.ukserp.ac.uk:UKLLC' is highlighted in yellow. Under 'Clone with HTTPS', the URL 'https://gitlab.ukllc.ukserp.ac.uk' is shown. Under 'Open in your IDE', there are links for 'Visual Studio Code (SSH)', 'Visual Studio Code (HTTPS)', 'IntelliJ IDEA (SSH)', and 'IntelliJ IDEA (HTTPS)'. The main content area shows the project's commit history with a table listing files and their last commit.

4. Open Git BASH and change directory ("cd #insertFilePath") to folder in your project file storage area where you want to clone files to and type: (replacing the repository name with the one copied in the previous step) :

```
tesri@ukllc-float-18 MINGW64 /s/LLC_9999/t4/tclone
$ git clone git@gitlab.ukllc.ukserp.ac.uk:UKLLC/ReadAll/testproj6.git
```

5. If successful you'll have a folder written to your chosen directory containing the repository contents.

4.6 Data refreshes

This section is forthcoming.

4.7 Reporting concerns about data

All researchers **must** contact the Data Team (support@ukllc.ac.uk) as soon as possible if they have **any concerns** that any of the datasets they have access to have NOT been reasonably de-identified or if there are **any concerns** about the quality of the data.

4.8 Detecting and preventing scope creep

All analyses performed by researchers MUST lie within the approved scope of the project – see Audits section below. Researchers should frequently revisit their approved application form and data request form, taking particular note of the following fields:

- Title
- Abstract
- Lay summary
- Project overview
- Project methodology
- Outputs
- Data requested
- Any imposed ethical or data restrictions.

Researchers should use their project documentation to determine if their analyses deviate from the approved scope or terms and conditions of the project. If there is **any doubt** regarding the permitted scope of your project, you **MUST** contact the Applications Team before proceeding any further (access@ukllc.ac.uk).

The Applications Team will determine whether your analyses or proposed analyses fall within the scope of your approved project, or whether you need to submit an amendment or a new application – see section 10. You will not be permitted to file-out data outputs that lie outside the scope of your approved project – see section 9.

4.9 Audits

Please be aware that all researchers agree to their **ONS Approved Researcher accreditation** and the **scope of their accredited projects** being **audited** and must provide full cooperation to internal UK LLC audit processes or audits being conducted by external organisations. All auditors commit to the confidentiality of the research project.

UK LLC Data Team staff will examine your syntax files and check that the variables you are accessing and including in functions, tabulations and visualisations align with your project documentation.

If any **potential non-compliances are identified**, the Data Team will contact the **lead researcher** on the project for further information and copy in the UK LLC Information Security Team (IST).

If any non-compliances have been identified, the IST will consider whether an Information Security Case Report should be completed. If appropriate, the IST will escalate non-compliances to the UK Statistics Authority's Research Accreditation Panel (RAP).

5. Using Software in the TRE

5.1 Software available

UK LLC supports researchers in using their preferred software package(s) and has a selection of software readily available. The following packages are pre-installed in the TRE and ready for use:

- SQL Management Studio
- Jupyter Notebooks
- Python
- R and R Studio
- Stata
- SPSS.

If you prefer to use other software, please contact support@ukllc.ac.uk and the UK LLC Data Team will be happy to assist you and to review your request. Please be aware that alternative software packages may incur a fee.

5.2 Using SQL Server Management Studio

SQL Server Management Studio can be used to view and query your data. To explore your database views, open the software package and connect to the server '**SERPSQL**'. Under 'Databases' in the Object Explorer, navigate to the '**UKSERPUKLLC**' folder, where you can explore your 'Tables' and 'Views'. All data and metadata can be found in 'Views'. Codelists can be found in 'Tables'.

In order to browse the data in a specific view, right click on the view name and 'Select Top 1000 Rows'. This will produce a grid-view of the first 1000 rows of data and a query window where data can be queried using SQL commands. This also works for tables.

5.3 Using Python

5.3.1 Background

If you are planning to use python and install any packages, you will need to create your own Anaconda environment. You should do this before running any code. This enables you to *pip* and *conda* install packages without barriers. We provide an automatic means of doing so, but you are welcome to create your own environment independently, installing your own subset of custom packages, should you wish.

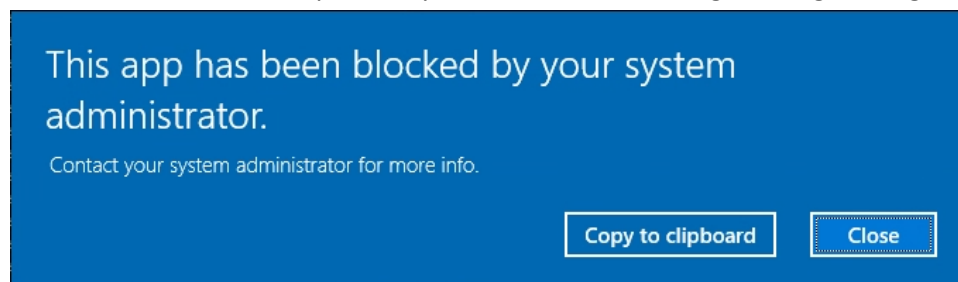
5.3.2 Creating a new environment

Alternatively, running "S:\ReadAll\python_help\env_helper\create_python_env.bat" will create a new environment as a clone of the base environment, meaning all standard packages included with the base Anaconda installation will be included. This is a more reliable approach if you are not experienced in creating environments and includes most commonly-used packages.

To run the batch file simply double-click on the file
"S:\ReadAll\python_help\env_helper\create_python_env.bat"

Things to be aware of whilst the environment is being created:

1. Creating a new environment can take around 2 hours. However, you can leave this running in the background.
2. The terminal may print "The following packages cannot be cloned out of the root environment:..." this is expected, and the environment should still create despite this message.
3. At the end of the process, you will see the following warning message appear:



This can be closed. It will appear 4 times. Your environment will still be successfully created despite the warning.

Once complete the script will have created a new environment in the location:
"P:/%USERNAME%/envs" called "env1". Therefore, the environment name will be
"P:/%USERNAME%/envs/env1".

The batch file ("create_python_env.bat") can be edited if you wish to store your environment in a different location or would like to name it differently. Please take a copy of the batch file and save in your project or personal area before editing. The environment will be created in the P: drive as this is a personal drive available only to you and always accessible from whichever remote desktop you are assigned on login.

5.3.3 Using your new environment

If you are using Jupyter Notebooks, Spyder or any other Anaconda tools for your analyses it is recommended that you launch these from Anaconda Prompt after activating your new environment. This will ensure you are running your new environment within the application. To do this:

- launch Anaconda Prompt
- type: conda activate "P:/%USERNAME%/envs/env1"

- replacing %USERNAME% with your actual username
- then within Anaconda Prompt type the name of the application you want to run: eg “spyder” or “jupyter notebook”

Note: for Jupyter notebooks you may want to change directory to your working directory before running the application.

We recommend using Jupyter notebook and we have prepared a template notebook with a comprehensive suite of helper functions. You will find this in:

“S:\ReadAll\python_help\notebook_helper”. There are two files of note here, “notebook_template.ipynb” and “notebook_helper.py”. Make sure to copy them both into your project’s working directory. The notebook template will guide you through how to pull your data and operate on it, using functions from the notebook helper program. You needn’t look into the underlying code in the notebook helper, but you are welcome to make changes should you wish.

5.4 Using R

If you are planning on using R, an R helper is provided in “S:\ReadAll\r_help”. These scripts are designed to help you extract your data from the database and to add variable and value labels using the *expss* R package. To use the helper:

1. Take copies of both scripts “r_label_helper.R” and “r_label_functions.R” and save copies to your working area. This is because will need to edit “r_label_helper.R” with your project specifics and requirements.
2. Once you have local copies, ensure both scripts are in the same directory as “r_label_helper.R” will expect to find “r_label_functions.R”

Open “r_label_helper.R” with RStudio and find comment “ACTION1: enter/update project number...”. Update the *proj_no* variable to your project number in form “LLC_0000”

```
#####
# ACTION1: enter/update project number in form "LLC_0001"
#####
proj_no = "LLC_0000"
```

3. Next run script up to and including:
and print
print(viewnames) #print list of views for inspection
4. You will see a list of your project views printed to the console and a “viewnames” variable will be created in your environment. Check “viewnames” are consistent with your expectation.
5. Next view the remainder of the script. Here there are 3 options are pre-scripted for retrieving data from the DB and adding the variable and value labels:
 - a. One table at a time
 - b. One study/data source at a time e.g. MCS
 - c. All tables at once (not recommended due to performance issues).

5.4.1 METHOD A – One table at a time

1. Under ACTION2a update the *data_source* and *table* name variables using lowercase e.g.:

METHOD A: label one table at a time
ACTION2a: define/update data_source AND table name to label (use lowercase)

data_source <- "genscot"
table <- "covidlife1_v0001_20211101"
2. Note your views are in the example form LLC_0001.GENSCOT_COVIDLIFE1_v0001_20211101

3. The *data_source* (usually study name) is in red, separated by the descriptive part (+versions and date) by an underscore. You will need to isolate these and assign to the *data_source* and *table* variables separately as shown above.

5.4.2 METHOD B – One study/data source at a time

1. Under ACTION2b update the *data_source* only using lowercase e.g.:

```
#####
# METHOD B: separate one study/source and label multiple tables at a time
# ACTION2b: define/update data_source to label (use lowercase)
#####
data_source <- "genscot"
```
2. This will retrieve all views from that data source and apply value and variable labelling. The output will be stored as a list of dataframes called "one_stud_dfs_w_labs"

5.4.3 METHOD C – Load all data at once

1. No update to script required here. Run code underneath:

```
#####
# METHOD C: label all tables at same time - WARNING - SLOW TO RETRIVE ALL DATA
#####
```
2. As stated in the comments this method is not recommended if you have large number of dataset as these will be slow to retrieve and add labels to.

You are welcome to modify the "r_label_helper.R" to suit your use case and use the example syntax to apply labelling to your own subsets of tables.

5.5 Using Stata

5.5.1 Using the helper to retrieve and label data

A Stata helper script is provided in "S:\ReadAll\stata_help". The script "py_stata_helper.py" is run by a batch file "run_stata_helper.bat". This program will extract data from your project views in the database, add variable and value labels and save stata format (.dta) files to your project working space in "S:\LLC_XXXX\data\stata_w_labs". To use the helper:

- You will first need to create your own python environment. See section 5.3.2 "Creating a new environment" for how to do this (if not already done)
- Once this program completes double click on "S:\ReadAll\stata_help\run_stata_helper.bat"
- The program will ask you to input your project number. Do this in the example form "0001"
- The command prompt will print to inform you when a dataset has been retrieved and whether value and variable labels have been found. The files will then be saved in .dta format in "S:\LLC_XXXX\data\stata_w_labs". The files can then be opened in Stata.

Warning: if your project has been provisioned a large number of views this process can take up to an hour to complete.

5.5.2 Manually retrieve and label data

To retrieve the full contents of a data view in Stata run:

```
odbc load, exec("select * from #INSERTVIEWNAME#") dsn(LLC_DB)
```

View names can be found in “available_views” folder in your project space. Replace #INSERTVIEWNAME# with the view name you’d like to retrieve this must include schema (usually your project number e.g.:

```
odbc load, exec("select * from LLC_9999.bcs70_basic_demographic_v0001_20211101") dsn(LLC_DB)
```

Variables and value labels can be retrieved via the following database views:

- DESCS.all_descriptions – for variable labels
- VALUE.all_values – for value labels.

5.5.3 Using Stata with Jupyter Notebook

If using Stata, we recommend building your analyses into a Jupyter Notebook. Jupyter Notebook allows you to execute code, create visualisations with narrative text in one document (a "notebook"), enabling researchers to easily share their code and research findings. Stata in Jupyter is invoked using the IPython kernel, so you can use Python and Stata in the same environment. To use Stata within Jupyter you will first need to create your own python environment. See section 5.3.2 “Creating a new environment” for how to do this. You will also need to run the Stata helper (see section 5.5.1) to convert your database views into Stata files. Once you’ve completed these 2 steps, you can begin with Stata in Jupyter...:

1. Launch Anaconda Prompt.
2. Type: `conda activate "P:/%USERNAME%/envs/env1"` replacing %USERNAME% with your actual username. This will switch you to your custom environment.
3. Change directory to where you want to save your workings by typing: `S:` and then `cd "LLC_9999\syntax"` replacing “LLC_9999\syntax” with target working directory.
4. Within Anaconda Prompt type: `jupyter notebook`. This will launch Jupyter Notebook with your custom environment and working directory pre-configured.
5. When Jupyter has loaded, select “New” and “Python 3” (Warning 404 message will flash up for approx. 10 secs – ignore):



6. You are now ready to execute code, in cell 1 type:

```
import stata_setup
stata_setup.config('C:/Program Files/Stata17', 'se')
```

This will configure your stata setup:

```
In [2]: import stata_setup
        stata_setup.config('C:/Program Files/Stata17', 'se')

STATA® 17.0
SE-Standard Edition

Statistics and Data Science Copyright 1985-2021 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC https://www.stata.com
979-696-4600 stata@stata.com

Stata license: 144-user network perpetual
Serial number: 401706344668
Licensed to: UKSeRP User
Swansea University

Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5,000; see help set_maxvars
```

7. You can then load a dataset and begin analyses. In each cell you intend to run stata code the first line must begin with `%%stata`. See example in below screenshot:

```
In [8]: %%stata
        use "S:\LLC_9999\data\stata_w_labs\GENSCOT_VACCINE_v0001_20211101.dta", clear
        ds

. use "S:\LLC_9999\data\stata_w_labs\GENSCOT_VACCINE_v0001_20211101.dta", clear

. ds
llc_9999_s~d cohort_desc days dt2 sex
age_band council_ar~e dose_number health_boa~e vacc_healt~e
avail_from~t data_source dt1 product_name
```

If you are new to Jupyter, it is recommended that you familiarise yourself using online tutorial/training material.

6. Core Data Files

The UK LLC make available a number of standard 'core' files with each project. An overview of these is given below.

6.1 CORE denominator file1

LLC_XXXX.CORE_denominator_file1_YYYYMMDD has one row per UK LLC participant, the cohort/study to which they belong and all UK LLC configurable permissions for the different domains of data. The date in the name of the dataset relates to the file1 (permissions) that your project is fixed to. Therefore participant permissions are correct as of this date. Permissions are updated by the study on a quarterly basis. This file allows users to:

- calculate linkage rates per study/cohort when joined to NHS D demographics and other NHS D datasets
- contextualise study/cohort participants within the UK LLC in relation to the study/cohort datasets.

6.2 CORE NHSD Presence

LLC_XXXX.CORE_NHSD_Presence_v0000_YYYYMMDD contains the number of appearances and the date of the most recent appearance for each participant for each available NHS data source in the UK LLC database. There is one row per participant. The table also includes a 'last seen date' for each participant as the most recent record across all NHS data sources. Some NHS data sources do not include record dates, therefore the presence table may include the counts of each participant's appearances in a data source without an associated date.

6.3 CORE NHSD Derived indicator

LLC_XXXX.CORE_derived_indicator_v0000_YYYYMMDD contains the most recent and most reliable record for certain key variables sourced from NHS data sources in the UK LLC database. There is one row per participant. The table sources data from the following tables and preferential uses data in this order:

1. Demographics
2. GDPPR
3. HESAPC
4. HESOP
5. HESAE

Currently included are the following variables:

- *sex*
- *Deceased*: from NHSD.mortality
- *Date of death*: from NHSD.mortality
- *ethnic*: NHS D coding system, see values table for coding lookup
- *dob_year_month*
- *last_seen_date*: last date record in any NHS D dataset

6.4 CORE NHSD LSOA11

LLC_XXXX.CORE_nhsd_lsoa11_v0000_YYYYMMDD has one row per health interaction where the project has selected the cohort/study and where participant permissions are in place. The indicator picks from the following tables:

- NHSD.GDPPR
- NHSD.HESAE
- NHSD.HESAPC
- NHSD.HESOP

These datasets contain lower super output area (LSOA) 2011. LSOA is encrypted in the dataset, because geographical units smaller than region are not permitted 'in-the-clear' in the UK LLC TRE. The dataset contains the following variables:

- *record_date*: date stamp from health record
- *lsoa11cd_e*: encrypted LSOA 2011 from health record
- *origin*: NHS D dataset the LSOA originated from

This dataset can be linked to **CORE_LSOA11_geo_indicators** (documented in next subsection) to add in geographical indicator variables associated with encrypted LSOA (*lsoa11cd_e*).

6.5 CORE LSOA11 geo indicators

This file is intended to be linked to **CORE_NHSD_LSOA11** to add the geographical indicators to the LSOA from the health record. The dataset contains the following variables:

- *imd2019_#subdomain#_q##*: IMD2019 with subdomains in deciles, quintiles, quartiles and tertiles. 1 is most deprived and ceiling value is least deprived
- *ctry17nm*: country name 2017
- *gor10nm*: region name 2010
- *people_km2_2020*: population density (number of people per Sq Km – mid 2020)

- *ruc11cd_v2*: urban rural classification collapsed from 8 to 5 categories
- *origin*: NHSD dataset LSOA originated from
- *record_date*: date health record was generated

To link LLC_XXXX.CORE_nhsd_lsoa11_v0000_YYYYMMDD with LLC_XXXX.CORE_lsoa11_geo_indicators_v0000_YYYYMMDD:

1. Retrieve data from database via helper syntax
2. Link datasets on *lsoa11cd_e* field. Example of stata syntax:

```
* load NHSD LSOA dataset
use "S:\LLC_9999\data\stata_w_labs\CORE_nhsd_lsoa11_v0001_20221217.dta", clear
* many to one merge on the lsoa11cd_e variable
merge m:1 lsoa11cd_e using
"S:\LLC_9999\data\stata_w_labs\CORE_lsoa11_geo_indicators_v0001_20220913.dta"
* drop LSOAs not linked to any participant health records
drop if _merge == 2
```

CORE_nhsd_lsoa11 is a long dataset typically with millions of rows, depending on size of data request. It is therefore recommended that you subset both or either of these datasets before linking/processing/saving. An example of this would be to select the quantile of IMD that you are going to use and keep these variables only. This will ensure the dataset size remains as manageable as possible.

Note: the geo indicators were previously provided as a pre-linked dataset:

CORE_NHSD_Geo_Indicator_v0000_YYYYMMDD. This has been deprecated because the file size was too large for some software programs with our desktop configuration, hence why it is now provided as 2 separate smaller files which can be subsetting and linked to suit use case.

6.6 NHSD DEMOGRAPHIC SUB

Where any NHS England data is selected, a 'DEMOGRAPHICS_SUB' block is provided. This block is a sub-block of the NHS D DEMOGRAPHICS dataset, which contains study ID, DOB (month/year) and gender

If an individual has a record in this tables it indicates a successful linkage has been established to the English NHS Patient Demographics Service by the UK LLC. This table gives a defacto denominator for participants for whom an NHS D linkage has been established (but does not necessarily mean the individual has any record in any of the NHS D data blocks).

NOTE: People may exist in this dataset twice if they are in multiple LPS. Currently, participants who are in multiple LPS cannot be linked between studies. However, this functionality has been factored into the design of the UK LLC and will be implemented in the near future.

7. Wellcome Trust COVID Questionnaire Lookup

Many LPS datasets within the UK LLC database contain questions adapted from the standardised Wellcome Trust COVID Questionnaire. This questionnaire includes generic COVID related questions. You can use this 'spine' of Wellcome Trust questions to find similar variables across different datasets in the UK LLC TRE.

The similarity of possible combinations of questionnaire question and dataset variables is stored in the "WELLCOMETRUST_COVID_Questionnaire_vXXXX_YYYYMMDD" table. You should use this table through the "Wellcome Tust COVID Questionnaire Lookup Workbook" Jupyter Notebook, available in the READALL folder of the S: drive at "S:\Readall\Wellcome Trust COVID Questionnaire lookup".

The notebook walks through the process of loading and viewing the lookup table, searching for similar variables and joining descriptive metadata to the search results.

To run the notebook, first copy the notebook from the READALL folder into your project's working directory. Run the notebook by navigating to your working directory in anaconda prompt and entering "Jupyter notebook". The notebook includes instructions for its own use.

8. Getting Files into the TRE

Throughout the duration of your project, you will likely need to use information (i.e. syntax, documentation) held outside the TRE, inside the TRE to facilitate your research. In order to get files into the UK LLC TRE, you must submit a ('File-In') request through the [SeRP Portal](#), which will be reviewed by the UK LLC Data Team. We aim to review all File-In requests within 1-2 working days. In order to get files into the TRE follow the SeRP UK guidance available here:

<https://docs.hiru.swan.ac.uk/display/HDK/File+In+User+Guide>

NOTE: Data **can not** be filed into the TRE by researchers. If your project requires data not currently available in the TRE this will need to be discussed with the UK LLC Applications Team and may be agreed on a case by case basis (access@ukllc.ac.uk).

9. Getting Files out of the TRE

9.1 Introduction

An output is any content you want to take out of the TRE – this includes data outputs, syntax and documentation. Before any output is released it **must be screened** to ensure it is not disclosive, i.e. safeguards the confidentiality of data subjects, and is within the scope of the project and data owner permissions.

All researchers should apply appropriate Statistical Disclosure Control (SDC) as explained in the **ONS Safe Researcher Training** that all researchers completed prior to being permitted access to the UK LLC TRE. Excellent additional resources are:

- The SDC Handbook produced by the Safe Data Access Professionals network: [SDC Handbook \(securedatagroup.org\)](#)
- SRS Output Checking Guidance Document Work strand: Statistical Disclosure Control produced by ONS: [Supporting your research project - Office for National Statistics \(ons.gov.uk\)](#).

9.2 File-out process

You must submit a file-out request through the [SeRP Portal](#). All requests are triaged by the UK LLC Data Team. All analytical/statistical outputs are reviewed independently by two members of the specialist SDC team based at Swansea University (SAIL Databank).

The **UK LLC Data Team** will review requests for file-outs containing syntax, documentation and codelists and will process requests within 1-2 working days.

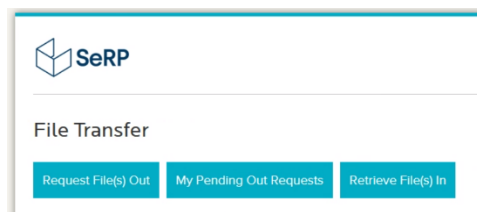
The **SDC Team at SAIL Databank** will review all analytical data-related outputs and carry out SDC before approval. As analytical/statistical outputs require a double review this can take several days to be approved depending on complexity.

To request files out of the TRE follow the steps below:

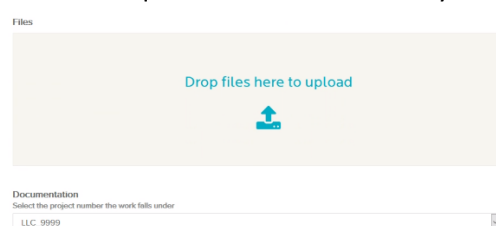
1. Launch 'Data Transfer (In-Out)' from the desktop (best launched in Firefox).



2. Select 'Request File(s) Out'



3. Request files out and select your project number (e.g. llc_9999).



9.3 File-out rules

It is a critical principle of the UK LLC TRE that only **anonymous population level aggregate data** (e.g. tables of results) that are **within the scope of your approved project can be filed out from the TRE**. It is not permitted to file out any individual level data from the TRE. All outputs must be assessed for disclosure risk and transformed to control for identified risks prior to being permitted to leave the TRE.

All users must agree to uphold these principles:

- All outputs must comply with the project's permissions and scope.
- **The SDC threshold is 10 data subjects.** Unweighted counts <10 and statistics (e.g. weighted counts, percentages, means, etc.) derived from groups containing <10 data subjects must be suppressed. Zeros are included in this threshold, i.e. counts ranging from zero to nine, inclusive, are not permitted.
- All users must use their knowledge from the **ONS Safe Researcher Training course** (and other relevant training) to make best efforts to ensure outputs are anonymous before submitting them for 'data out' checks.
- The findings of the file-out review are binding and research users must correct any identified issues.

Failure to adhere to these conditions may result in sanctions (such as users being required to undergo renewed training).

10. Amendments to Projects

If a researcher wishes to amend their approved project after their data have been provisioned in the TRE, they must submit an **amendment**. Amendments fall into the following five categories:

1. Requesting new data.
2. Adding/removing a researcher.

3. Moving organisation.
4. Extending TRE access period.
5. Changing research question.

Amendment forms are available on the [Resources for Researchers SharePoint](#) and should be submitted to the Applications Team for processing (access@ukllc.ac.uk).

11. Finishing a Project

11.1 Request final file outs

Request the final file-outs that are required for your publications/presentations.

11.2 Share your syntax

It is expected that syntax and documentation associated with your analyses are returned for re-integration into the resource. This is detailed in section 4.5 of this user guide. It involves making a file-out request and marking in the comments that it is “For sharing via GitLab/GitHub”. The contents will be disclosure checked. If checks are passed the contents will be pushed to the shared subgroup in GitLab called “ReadAll”. Researchers must also push the approved files to their project repository on the external public GitHub.

11.3 Share your derived data outputs

To return data outputs for re-integration into the UK LLC resource, email support@ukllc.ac.uk with the TRE file path of the data file(s) for reintegration. This email must also supply the project name and GitLab repository containing syntax used to create the data and associated data documentation. We cannot process derived data without syntax and documentation. Please ensure that the data are named following the [UK LLC Naming of Projects and Data-related Outputs Policy](#).

Following a disclosure risk assessment, data files returned will be written as a database table to the UK LLC database and will be available to requests by other research projects. The data file(s) should be in stata (.dta) or .csv format and have variable and value labels where appropriate. For stata these can be embedded in the .dta file. For .csv, these should be supplied as 3 separate files. Here is an example of the naming convention for csv files (stata should follow the data convention, adjusting file extension):

Data: LLC_9999_asthma_diag_data_v0001_20221208.csv

Variable labels: LLC_9999_asthma_diag_data_v0001_description_20221208.csv

Value labels: LLC_9999_asthma_diag_data_v0001_values_20221208.csv

12. Publishing or Presenting your Work

The [UK LLC Publication Policy](#) describes the requirements for researchers when publishing **papers and similar outputs** based on data accessed in the UK LLC TRE. All such outputs must be submitted to the UK LLC (access@ukllc.ac.uk), along with the [Publication Checklist](#), for review prior to submission.

The UK LLC does not require to see **presentations** before they are delivered, but researchers must acknowledge the use and funding of the UK LLC’s resource and the collaborative contribution of the LPS and other data owners and partners – for acknowledgements, logos and other useful materials see the [UK LLC GitHub repository](#).

13. User Support

The UK LLC Data Team maintains a library of self-service user help tools in multiple formats, including a user guide, helper syntax and [YouTube videos](#). In addition, all active users are invited to join the monthly **UK LLC Analyst Group meetings** and the **UK LLC Data Users' Google Group**.

If users are unable to find the help they require, the UK LLC Data Team can be contacted at support@ukllc.ac.uk. The UK LLC Data Team uses a JIRA ticketing system to track requests. You will receive an email notification when the request is received and as the ticket changes status until it moves to done/complete. Progress can be tracked and comments can be added.

If users experience issues with the software or the SeRP environment they should contact helpdesk@chi.swan.ac.uk in the first instance.

13.1 UK LLC Data Users' Google Group

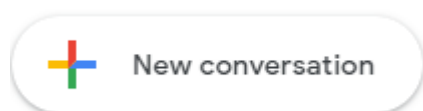
The group should be used to discuss UK LLC data-related insights and issues with fellow users.

To join the Google Group:

1. Go to this link: <https://groups.google.com/u/2/g/uk-llc-data-users>.
2. Click "Ask to join group" button.
3. Await approval.

Basics on how to use:

1. To start a new topic click:



2. To reply to an existing conversation, click on "Reply all", insert message and click "Post message"
3. To set your email notification settings go to "My membership settings" from left pane.

14. Making a complaint

UK LLC aims to deliver a high quality and inclusive service that empowers researchers to conduct impactful research in the public interest.

If you feel the service you have received from UK LLC has fallen below your expectations, you can file a complaint.

If you are accessing **LPS data and health data**, you should email **UK LLC directly**: info@ukllc.ac.uk

If you are also accessing **government administrative data** (DWP, DoE, HMRC) through the Digital Economy Act (DEA) legal gateway, you may alternatively choose to email the **UK Statistics Authority** (the body that accredits UK LLC as a DEA Accredited Processing Environment): Research.Accreditation@statistics.gov.uk.