



UK Longitudinal Linkage Collaboration
Population Health Sciences
Bristol Medical School
Canynges Hall
39 Whatley Road
Bristol BS8 2PS

The UK Longitudinal Linkage Collaboration (UK LLC)

File 1 Checker User Guide

PUBLIC
Version 2.0
2nd November 2022

1. Introduction

The File checker is a user tool for verifying the contents of a File 1 submission are in line with the requirements set out in “UK LLC File Formatting Guidance for Depositing Data in the UK LLC Resource”. The checker verifies that field names and values are of expected syntax and data type. It makes no judgement on the contents of the files other than their legality under formatting rules. It is very important that File 1s are cleared through the checker to avoid complications loading them into the UK LLC databank.

The checker includes a section for File 1 Documentation that should be sent to the UK LLC (support@ukllc.ac.uk). Information is automatically loaded from a File 1 where possible and cross referenced for correctness. The documentation is saved in a standard JSON format ready to be sent to the UK LLC.

The program is available as a graphical user interface that can be run as an executable or with python. The code is available at <https://github.com/UKLLC/File-Checker>.

2. Installation

Go to the UK LLC GitHub at <https://github.com/UKLLC/File-Checker> and download the code as a zip file. Extract the contents into its own folder. The checker will produce output files in the ‘outputs’ folder. Please make sure the directory has read and write permissions for this to work. When extracted, the directory should contain the files shown below

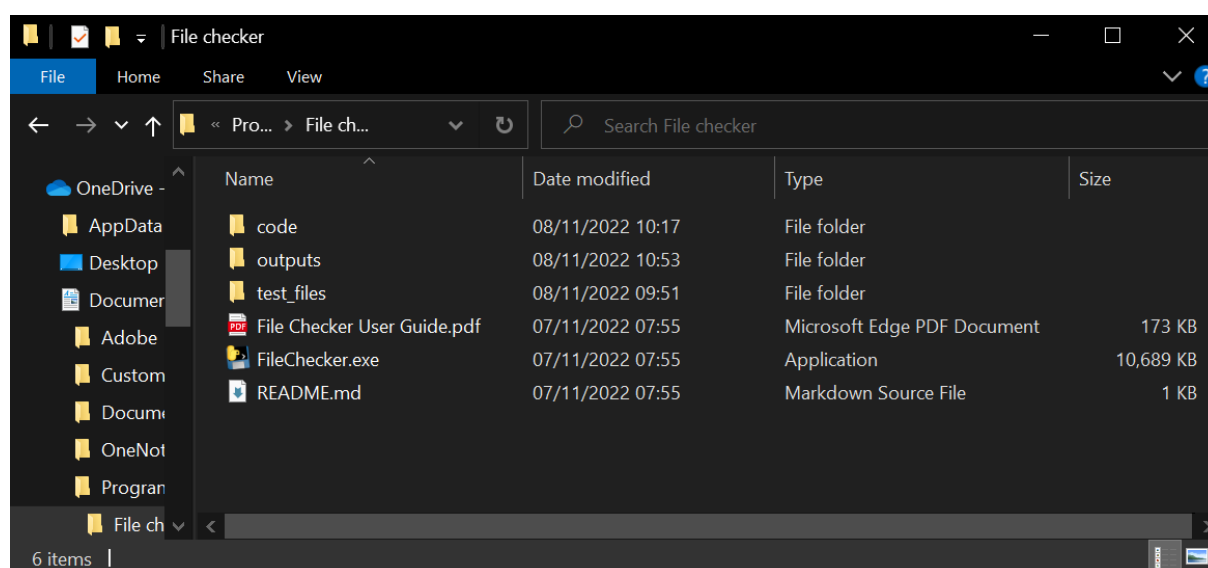


Figure 1. File Checker Directory

3. Running

Run the checker program by double clicking, or right clicking and selecting “run”, the FileChecker.exe application. This will launch a graphical user interface. If you are unable to run executable files, you can instead run the application with Python. You will need an installation of Python 3. Open a command prompt window and navigate to the code folder in the checker’s directory. Run the application by typing ‘python ui.py’.

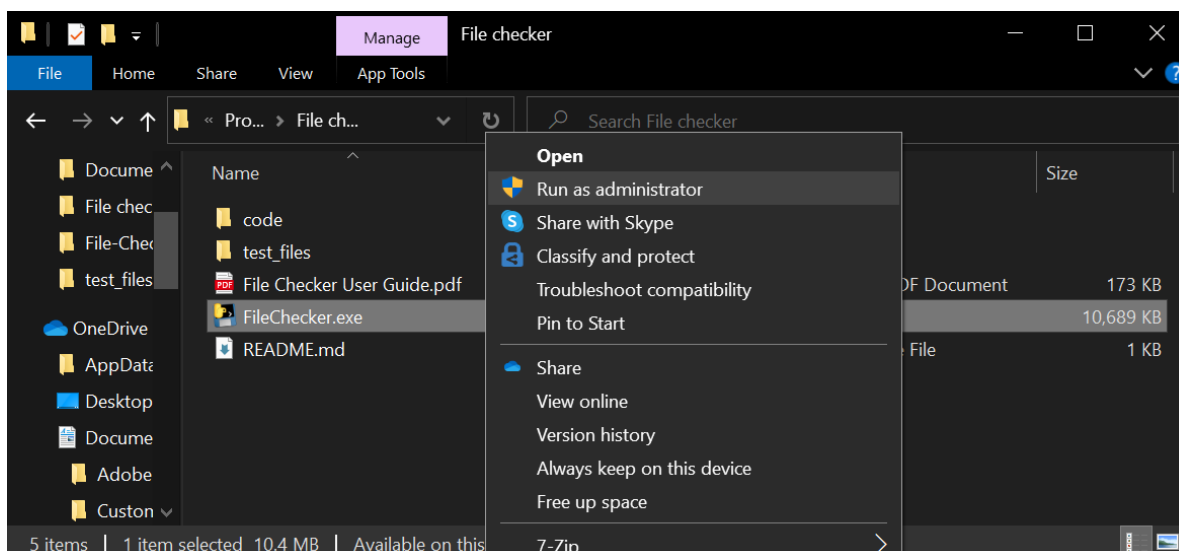


Figure 2. Running with FileChecker.exe

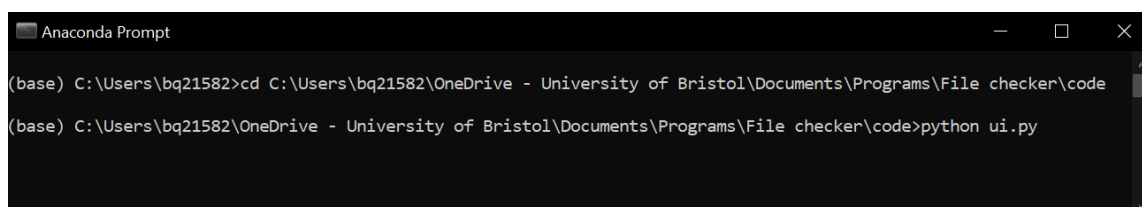


Figure 3. Running with Python

When the interface appears, start by loading a File 1 csv by clicking the 'Load File 1' button. To begin the file checking process, click the 'Start' button. The checks might take several minutes depending on the size of the file. The results of the file checker will be presented in the text field directly below the start button and saved as a text file in the outputs folder in the format: '[input_file_name]_Output_Log_[time:HHMMSS].txt'.

File 1 Integrity Checks

Please select your File 1. The file must be in CSV format.

Load File 1

Click 'Start' to begin automated file 1 integrity checks.
Please wait until the automated checks are completed before filling out the File 1 documentation section

Start

Integrity checks output:

File 1 Documentation

Please make certain the number of participants included in the sample and excluded from the sample add up to the total number of participants in the cohort. If you are unable to categorise exclusions, please include them in field 8: 'other'.

Date: 08/11/2022

File name:

Study name:

Figure 4. File 1 Integrity Checker UI

The output is a list of all errors encountered, including a small description of the error and lines of the file where it was encountered where appropriate. If more than 10 lines suffer from the same error, only the first 10 lines will be listed for the sake of brevity, with the understanding that the error is likely widespread throughout the file.

File 1 Integrity Checks

Please select your File 1. The file must be in CSV format.

Load File 1

loaded 'C:/Users/bq21582/.../test_files/general_bad.csv'

Click 'Start' to begin automated file 1 integrity checks.
Please wait until the automated checks are completed before filling out the File 1 documentation section

Start

Checks completed.

Integrity checks output:

```

-----
Date Format Error
Invalid format for field ADDRESS_START_DATE. Date should be in the format DD/MM/YYYY
Line(s) 3
-----
Value Error
Invalid value for field STUDY_ID. Please refer to the specification for correct field formatting guidelines.
Line(s) 5, 6
-----
Value Error
Invalid value for field ROW_STATUS. Please refer to the specification for correct field formatting guidelines.
Line(s) 3, 4, 6

```

File 1 Documentation

Please make certain the number of participants included in the sample and excluded from the sample add up to the total number of participants in the cohort. If you are unable to categorise exclusions, please include them in field 8: 'other'.

Date: 08/11/2022


File name: general_bad.csv

Study name:

Figure 5. File Checker Example

After the automated file checks are completed, fill out the 'File 1 Documentation' section. The date, file name, study name, row count and included participants will be automatically filled from the loaded file where possible. Make sure the number of exclusions and inclusions add up to the total number of participants in the study.

When finished, click 'Save & submit'. This will save the File 1 Documentation inputs as a JSON file in the 'outputs' folder in the format: 'File1_Doc_[filename].json'.


UKLLC File 1 Checker and Documentation

File 1 Documentation

Please make certain the number of participants included in the sample and excluded from the sample add up to the total number of participants in the cohort. If you are unable to categorise exclusions, please include them in field 8: 'other'.

Date:	08/11/2022
File name:	general_bad.csv
Study name:	Demo
Row count:	6
Expected date File 1 uploaded to DHCW:	08/11/2022
1. Please enter the total number of participants (n) in the cohort (enrolled sample/headline denominator)	10
2. Please enter the number of participants (n) excluded because they died on or before 31/12/2019	2
<p>i.e. participants who died and whose death is not likely to be related to COVID 19. We would expect this number to be 0 because these participants can have their data flow to the UK LLC, unless there is specific study policy precluding them.</p>	
3. Please enter the number of participants (n) excluded because they died on or after 01/01/2020	2
<p>It is essential that data for participants who have died during the COVID 19 pandemic (on or after 01/01/2020) continue to flow to the UK LLC TRE, unless this directly violates Study policy. Therefore, we would expect this number to be 0.</p>	
4. Please enter the number of participants (n) excluded because they have withdrawn from the LPS	2
5. Please enter the number of participants (n) excluded because they have specifically dissented to the use of their data in the UK LLC TRE	0
6. Please enter the number of participants (n) excluded because they have dissented to record linkage (i.e. NHS Digital)	0
<p>While it is up to LPS whether they send data for participants who have dissented to record linkage (i.e. NHS Digital), please be aware that these participants can be sent to UK LLC with permissions set accordingly. Dissenting to record linkage does not preclude participants from the UK LLC resource, where study-collected data can be provided.</p>	
7. Please enter the number of participants (n) excluded because appropriate governance has not been established	0
8. Please enter the number of participants (n) excluded for 'other' reasons	3
9. The number of participants (n) included in the sample uploaded to NHS DHCW (i.e the number in your File 1 where UK LLC status (UKLLC_STATUS) is equal to 1 and Row_Status is equal to 'C')	1

Warning: automated checks detected problems with the file.
Only save and submit if you are certain the file is correct.

Save & submit

Figure 6. File 1 Documentation

You will receive an error message if your inputs are not in the expected format (text in a field requiring numbers). You must resolve this error before you can save the documentation.

You will receive a warning message if the 'Study name' entry does not match a recognised UK LLC study ID. Please make sure you are using the agreed upon ID for your study. If you are certain the ID is correct and you are still getting a warning message, you are welcome to continue with the save. In this case, please inform the UK LLC of this difficulty when you send the documentation.

You will also receive a warning message if the number of inclusion and exclusions do not sum to the total number of participants in the cohort. Please fill in the exclusions in as much detail as possible. If you are unable to detail the reason for a participant's exclusion, please add them to field 8, 'other'.

4. Outputs

When you have finished running the file checker, you should have at least two output files. The output log text files are records of the automated file checker's findings for your own reference. The File1_Doc JSON file should be sent directly to the UK LLC at support@ukllc.ac.uk. As always, please do **not** send the file 1 itself to the UK LLC.