

SLS UNDERTAKING FORM - (SYNTHETIC DATA)

An SLS Undertaking Form (Synthetic Data) must be signed by the researcher and any collaborators with whom they are working. The form covers storage, confidentiality and validity issues relating to your project to ensure you are aware of your responsibilities and the conditions you must conform to when using synthetic SLS data.

A synthetic version of your project extract will not be produced until a SLS Undertaking Form (Synthetic Data) has been signed by ALL members of the research team.

**Your Privacy**

The information you provide on this form will be treated in the strictest confidence. An electronic copy of this form will be securely stored with NRS and the SLS-DSU. A hard copy will be securely stored in the SLS-DSU. For further information on how your data is used, see:

<https://www.nrscotland.gov.uk/record-keeping/legislation/primary-information-legislation/data-protection>

<https://sls.lscs.ac.uk/more/privacy-statement/>

**Application details**

***Study title:***

***SLS study number***(provided by your Support Officer)

## Head of proposed study

|  |  |
| --- | --- |
| Title:        | Name:       |
| Position:       |
| Organisation:       |
| Address:       |
| Telephone:        | Fax:       |
| Email:        |

**Undertaking**

**Background**

The Scottish Longitudinal Study is now offering the option of receiving a synthetic version of a project extract. The provision of bespoke synthetic data will allow researchers to develop and refine their syntax and models using the synthetic data outwith the SLS Safe Setting before bringing these in to run on the real data.

Although these datasets are fully synthetic and contain no personal information, it is still important to treat them with care and caution. All synthetic data sets being transferred outwith the SLS network boundary should be treated as if they are data sets containing real data.

The controls and conditions associated with receiving and using SLS Synthetic Data are set out in this document. Before receiving a SLS synthetic dataset ALL members of the research team must have read and signed this agreement.

**Disclosure of information on individuals**

I understand that synthetic data does not contain personal information and as such I will not claim to have obtained or derived such information.

**Training**

I understand that all researchers will be required to have undergone the Safe Researcher Training and that certificated evidence of satisfactory completion of this training must be provided to the SLS Project Manager.

**Transfer of Synthetic Data**

I understand that any synthetic data files must be encrypted prior to transmission via e mail or media such as USB drive or DVD.

**Storage of Synthetic Data**

I understand that the synthetic dataset and resultant outputs should only be stored on an institution/work PC which has endpoint security measures installed[[1]](#footnote-1).

I understand that the synthetic dataset and resultant outputs should be stored securely and not on a shared drive. The synthetic dataset must only be accessible to the named research team (all of whom should have signed the Synthetic data undertaking form).

I understand that the file name of the synthetic dataset must not be changed and if subsets of the data are produced the filename must always contain the phrase ‘FALSE DATA’

I understand that the variable ‘false\_data’ must not be removed from the dataset and if any subsets of the data are produced they must always contain this variable

**Validity**

I understand that there is no guarantee of the validity of the SLS synthetic data, therefore results produced from analysis of the synthetic dataset must not be published.

I understand that I will be expected to provide feedback on the validity of the synthetic data compared to the real data to the SLS team.

**Outputs**

I understand that outputs from the synthetic dataset must only be viewed by the named research team (all of whom should have signed the Synthetic data undertaking form).

I understand that analysis and outputs from the synthetic data must not be published and that any analysis which I wish to publish must be run on the real data within the SLS Safe Setting.

**Retention of synthetic dataset**

I understand that I am required to destroy all copies of the SLS synthetic dataset and any outputs from the synthetic dataset following the completion of the SLS project.

**Breaches**

I understand that any breach of the conditions of the agreement must be reported immediately to the NRS SLS Project Manager.

**Note: All persons who will be using the data and are involved in the project must sign this undertaking.**

I have read and agree to the conditions above relating to the use of synthetic SLS data and recognise that any breach of these conditions may result in suspension of access to SLS data.

I will seek the agreement of the SLS Research Board for any extension of the project in terms of substance or duration.

|  |  |  |
| --- | --- | --- |
| Name of researcher(s)  | Signature | Date |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

A signed hard copy of this form should be posted to:



SLS-DSU

Longitudinal Studies Centre – Scotland (LSCS)

Room 1G1

Ladywell House

Ladywell Road

Edinburgh

EH12 7TF

Annex A

**What are synthetic data?**

Synthetic data are microdata records created to improve data accessibility whilst preventing disclosure of confidential information. They are produced by fitting statistical models to the original data and generating the synthetic data from these models – thus no records in the synthetic data correspond to real individuals.

The synthetic data set looks similar and behaves similarly to the real data, although the synthesising models may not capture all the relationships present in the real data. Usually the synthetic data will lead to the same conclusions as would be found from the real data, but this must be checked by running the final analyses on the real data.

To be absolutely clear, there are no records in the synthetic data set that are real. No real people can be identified from this data.

**How reliable are results from synthetic data?**

The synthetic spine data was produced with less emphasis on validity of the data and more emphasis on providing a resource which is freely available and can be used as a training dataset. All transitions will be accurate when aggregated to age (group), although not necessarily when aggregated to another variable.

* **We encourage researchers to** **use this dataset as a training or preparation resource only**.

In contrast, the bespoke synthetic datasets do aim to preserve the relationships between variables that exist in the real data as much as possible in the synthetic data.  The SLS synthetic dataset will look and behave relatively similarly to the real data. However, there is no guarantee of the validity of the results obtained from the SLS synthetic data.

* **The results produced from analysis of the synthetic dataset** **must not be published**.
* **Any analysis for publication must be run on the real dataset** **within the SLS Safe Setting and all outputs from this must be disclosure checked as standard.**

All researchers using synthetic data will be expected to liaise with their SLS Support Officers on how results for the real and synthetic data compare, so that this can be reported to our development team.  This feedback will be used in any future developments of the product.

**Why use synthetic data?**

Due to the confidential nature of the data held within the SLS, the data can only be accessed within the SLS Safe Setting of the National Records of Scotland (NRS) office in Edinburgh.  As synthetic data do not include records from real individuals they can be made available outside the SLS Safe Setting.

The synthetic spine is intended to provide a taster of the longitudinal aspect of the SLS data by providing five variables at two time points (2001 and 2011).

Bespoke synthetic datasets enable users to conduct preliminary analysis outside the safe setting.  The synthetic datasets can be used to develop analysis code that will be run on the real data, either when the user visits the SLS Safe Setting or by the staff of the SLS-DSU.

Synthetic data are valuable for preparing syntax and for training purposes however synthetic data cannot be used for final analysis as they are not real data.

**How to access SLS synthetic data**

The synthetic SLS spine data set is available [**https://calls.ac.uk/guides-resources/synthetic-ls-data/**](https://calls.ac.uk/guides-resources/synthetic-ls-data/)

SLS bespoke synthetic datasets can be requested during the SLS application procedure and will be created once the project has been approved by the SLS Research Board.

Before receiving the SLS bespoke synthetic dataset the research team will be asked to sign the SLS Undertaking Form (Synthetic Data).

* **The dataset will not be released until this has been signed by ALL members of the research team**

Once all the approvals are in place the dataset will be emailed to the lead researcher in an encrypted format.

The creation of good quality synthetic data can be a complex task and this may restrict what and when we can provide.  The following limitations apply:

* Priority will usually be given to users who cannot easily visit the safe setting.
* If your extract has many variables we may ask you to select a subset for your synthetic data (i.e. do some initial data preparation/management deriving variables for the synthesis).
* SLS person and family identifiers will not be included in the synthetic data.
* If your extract includes several files you will be linking together via identifiers, then you will not at present be able to link the synthetic data.  However, once you have linked the real data a synthetic version of the linked data can be provided.
* We now have general permission from PBPP to provide synthetic versions of NHS data, such as hospital admissions.  However, intention to create a synthetic extract should be highlighted in the PBPP form.

**How are synthetic data created?**

The two synthetic products that are available from the SLS are produced in different ways:

**1. Synthetic spine data**

The SLS ‘spine’ dataset is generated using the 2011 Scotland’s Census Teaching File dataset available from the National Records of Scotland:

<http://www.scotlandscensus.gov.uk/microdata-teaching-file>

and a series of 2001 to 2011 transitional probabilities of key demographic variables taken from the SLS.

The variables included are:

* Age (10 year groups)
* Marital Status
* General Health
* Religion
* Approximated Social Grade

A series of algorithms are used firstly to estimate the numbers of individuals in a particular age group undergoing each longitudinal state transition (eg. Never married in 2001 to Married in 2011 or Good health in 2001 to Good health in 2011) and then allocate these changes (or not) to the appropriate number of individuals in the Census dataset, resulting in a new, plausible, SLS-like dataset which will include data from both 2001 (synthetic) and 2011 (real) for all individuals.

For more detailed information see  ‘A Synthetic Longitudinal Study for the United Kingdom’

<https://calls.ac.uk/wp-content/uploads/NewSpineWriteUpV2.0.pdf>

**2. Bespoke synthetic extracts**

Bespoke synthetic extracts are produced using the R package synthpop in response to user requests:

<https://cran.r-project.org/web/packages/synthpop/index.html>

Variables are synthesised one by one using sequential regression modelling.  This means that each synthetic variable is modelled separately and this variable’s relationship to all other variables in the real dataset is taken into account.  This ensures that when analysis of the full dataset is performed the researcher will get results which will usually be very similar to results if this analysis was performed on the real data.

The synthetic data are produced from the user’s extract by staff at the SLS-SU.  This can be a complex task and users are expected to work with staff to facilitate their work.

1. Endpoint security is a method of protecting a corporate network. This is usually achieved by monitoring the status, activities, software, authorisation and authentication of all endpoint devices (PC, laptop, tablet etc.). Security software should be installed on any endpoint device, as well as network servers. Such software may include antivirus, antispyware, firewall and a host intrusion prevention system.

 [↑](#footnote-ref-1)