International Journal of Population Data Science

Journal Website: www.ijpds.org

Four checks for low-fidelity synthetic data: recommendations for disclosure control and quality evaluation

Gillian M. Raab[1,*], Sophie McCall[2], and Liam Cavin[3]

Submission History

Submitted:	24/01/2025
Accepted:	11/07/2025
Published:	28/10/2025

[1] Scottish Centre for Administrative Data Research, Futures Institute, 1 Lauriston Pl, Edinburgh, EH3 9EF
[2] Research Data Scotland, 1 Lauriston Pl, Edinburgh, EH3 9EF
[3] National Records of Scotland, New Register House, 3 West Register Street, Edinburgh, EH1 3YT

Abstract

Confidential administrative data is usually only available to researchers within a Trusted Research Environment (TRE). Recently, some UK groups have proposed that low-fidelity synthetic data (LFSD) be made available to researchers outside the TRE, to allow code-testing and data discovery. There is a need for transparency so that those who access LFSD know how it has been created and what to expect from it.

Relationships between variables are not maintained in LFSD, but a real or apparent data breach can occur from its release. To be useful to researchers for preliminary analyses LFSD needs to meet some minimum quality standards. Researchers who will use the LFSD need to have details of how it compares with the data they will access in the TRE clearly explained and documented.

We propose that these checks should be run by data controllers before releasing LFSD to ensure it is well documented, useful and non-disclosive.

1 **Labelling** To avoid an apparent data breach, steps must be taken to ensure that the synthetic data (SD) is clearly identified as not being real data.

2 **Disclosure** The LFSD should undergo disclosure risk evaluation as described below and any risks identified should be mitigated.

3 **Structure** The structure of the SD should be as similar as possible to the TRE data.

4 **Documentation** Differences in the structure of the SD compared to data in the TRE must be documented, and the way(s) that analyses of the SD expect to differ from those of data in the TRE must be clarified.

We propose details of each of these below; but a strict, rule-based approach should not be used. Instead, the data holders should modify the rules to take account of the type of information that may be disclosed and the circumstances of the data release (to whom and under what conditions).

keywords
synthetic data; low fidelity; disclosure risk; administrative data; transparency

*Corresponding Author:
Email Address: gillian.raab@ed.ac.uk (Gillian M. Raab)

https://doi.org/10.23889/ijpds.v6i1.2972

Introduction and background

In the UK and elsewhere there is increased interest in using administrative data for research to inform policy in area such as health, education and social policy [1, 2]. Several UK organisations have been established to facilitate making data from administrative sources available to researchers. They also support the linkage of such administrative data to other sources, such as Censuses or government surveys. The Economic and Social Research Council funds ADR UK (Administrative Data Research UK)[1] and its partner organisations in England, Scotland, Wales and Northern Ireland[2] to work with Government departments and National Statistics agencies to make administrative data available to researchers. HDR UK[3] performs a similar function for health data and Research Data Scotland[4] integrates data from many sources in Scotland. Researchers must apply for access to the data, usually available in a Trusted Research Environment (TRE). The process of applying and gaining access to the data in the TRE can be lengthy. The provision of synthetic data to researchers can allow them to develop their analyses and thus shorten the time between completing an application and obtaining results. The benefit of making synthetic data available is even greater where a visit to a safe setting is required to access the TRE. This is the case for the Scottish Longitudinal Study that provides a synthetic data option for users [3, 4]. Van Kestern [5] has argued that synthetic data is required to democratise research with sensitive data.

Methods for creating synthetic data (SD) for disclosure control have developed over the 30 years since this was first proposed [6], as described in two recent reviews [7, 8]. Synthetic datasets can be created using information from the original data, usually held within the TRE, for several purposes.

> **Purpose 1**: to allow users to discover the features of the data in the TRE and to develop code that will then be run on the data available in the TRE.
>
> **Purpose 2:** to allow users to develop analysis plans by using relationships in the synthetic data that will be expected to approximate those in found from the TRE data
>
> **Purpose 3:** to train researchers to understand the methodologies they will need to use data sets that are only available in TREs [9].

Kokosi et al [10] provide an overview of synthetic administrative data for research and propose terminology, initially suggested by the UK Office of National Statistics (ONS) [11]. They categorise synthetic data on a spectrum from *low fidelity* (minimally disclosive, minimal analytic value) to *high fidelity* (more disclosive, more analytic value)[5]. Purpose 1 can be achieved by SD at the low end of the spectrum. Purposes 2 requires the highest-possible fidelity, since preliminary analyses can influence the final results of an investigation. Purpose 3 requires at least moderate fidelity to ensure the training data appears realistic.

It is difficult to produce high fidelity SD that will reproduce all the complex relationships between variables in large administrative data bases. There is also concern about possible disclosure risk from SD that matches the original data too well. Although there has been some recent work on methods of assessing disclosure risk from SD [12, 13], experience with these metrics is limited. Thus several UK organisations have suggested that the priority for holders of sensitive administrative data should be to produce low-fidelity synthetic data for Purpose 1 [14–17].

As Bharat et al. [18] point out, the terminology used for SD applications is not standardised and can be confusing. Here we are using the term low-fidelity synthetic data (LFSD) for SD created by methods that do not attempt to model the relationships between variables. We know of two examples that currently make LFSD, created from UK administrative data, available to researchers. NHS England's *Artificial Data Pilot* [17] makes LFSD versions of several types of Hospital Episodes freely available to download. The code used to produce these SD episodes is available via GitHub[6]. Research Data Scotland offer LFSD versions of their data holdings, including those created from the 2001 and 2011 Censuses of Scotland [16]. In this paper we review methods for creating LFSD and illustrate them in the Appendix with a case study based on the 1901 Census of Scotland.

As the two reviews cited above [7, 8] point out, there are many methods that have been proposed and inplemented to evaluate the utility of synthetic data, but very few practical methods for evaluating disclosure risk. LFSD is a special case where procedures for evaluating disclosure risk may be easier to define. LFSD needs to be evaluated for its usefulness and disclosure risk. The goal of this short paper is to outline the steps that a data custodian should take to ensure that their LFSD is useful and is unlikely to pose a potential disclosure risk.

Steps in preparing synthetic data

Figure 1 illustrates the steps used in making data available for researchers. A raw data set is produced at step 1 either by downloading from an administrative data system or from survey or census data. The raw data may be from a single source, or from linkage of different sources.

At step 2 the data is prepared to make it suitable for analysis. This may involve steps to reduce disclosure risk (anonymisation and suppression of potentially disclosive values) and to improve its usefulness to researchers (correction of inconsistent values and imputation of missing values or missing records). At this step a research-ready data set is created. A review by Garth-Lone et al. [19] identifies the characteristics that make administrative data research-ready. They emphasise that the data needs to be curated to ensure its quality and documented to ensure transparency. Research-ready data may be used for internal analyses or supplied to external researchers either as an open public-use file (PUF) file[7], or for restricted use by researchers to analyse on their own

[1] Administrative Data Research UK https://www.adruk.org/.
[2] https://www.adruk.org/about-us/our-partnership/.
[3] Health Data Research UK https://www.hdruk.ac.uk/.
[4] Research Data Scotland https://www.researchdata.scot/.
[5] See Table 1 in the paper and an extended version in the supplementary material.
[6] See https://github.com/NHSDigital/artificial-data-generator.
[7] Another way that open data is supplied is the data set that sits behind a flexible table builder that allows external users to create their own tables (Thompson et al. 2013).

machines or to access in a TRE. Procedures to limit disclosure control at step 2 will depend crucially on the setting in which the data will be released [20]. These can include statistical disclosure control (SDC) procedures such as rounding and top- or bottom-coding for numerical variables, aggregation of categories and record swapping [21, 22].

At the final step the data in the TRE is used to create SD set(s) from the data created at step 2. SDC procedures can also be applied to SD before it is released to researchers and, prior to release, SD should undergo a series of checks. The proposals, outlined here, are appropriate for LFSD created by methods described below.

Methods for creating LFSD

Different methods can be employed at the low end of the SD spectrum. Here we will consider LFSD methods that do not attempt to maintain the relationships between variables. They use only data from the univariate distribution of each variable, but they vary on the detail provided about their distributions. These two examples are, respectively, the lowest and highest fidelity for LFSD:

1. By creating columns of data that match the structure and range of values in the original, but use only the metadata supplied and publicly available to users of the TRE. This is usually just a list of codes for categorical variables and either the range of values or a high and a low percentile for numeric data. This method does not require access to the data in the TRE. We will refer such methods as from codes.

2. When the exact marginal distribution of each variable is provided. This is equivalent to taking independent samples, with replacement, from each column of the original data. We will refer to this method as from exact marginals.

In the Appendix we present a case study where these two methods are used to synthesise data from the 1901 Census of Scotland. Method 1 will be expected a-priori to have very little, if any, disclosure risk because so little information about units in the original is used to create the SD. This expectation is fulfilled by the results from the case study.

There is a range of other possibilities between these two extremes that depend on the detail provided on the distribution of each variable. For example, the synthetic Hospital Episode data [17] creates aggregated anonymous summaries for each variable, which are then reviewed before being used to create the LFSD, but are not publicly available. In all cases the data used to create the SD may be altered to comply with privacy restrictions. For example, smoothing of numeric variables and other methods of noise infusion could be applied before the data is synthesised.

The sample size of the SD need not be the same as the original. Small sample sizes for SD may reduce disclosure risk by making it clear to the user that the data is not the original, but it is not clear how this will affect any formal measures of identity disclosure. This is discussed below and explored in the case study in the Appendix.

Even within these relatively simple methods, the usefulness of the SD can be enhanced by anticipating how the data will be used. For example, if medical data consist of dates of diagnosis and dates of death, then people in LFSD might be diagnosed after they die. To avoid this the data should be transformed into date of diagnosis followed by length of survival before the SD sets are created.

Disclosure risk (DR)

Data controllers have a responsibility to the people or other units whose data they hold. This includes ensuring that information about an individual unit is not made available to people who are not entitled to know it. Such an incident is referred to as a data breach. A data breach can have adverse consequences for the unit and may lead to loss of reputation and of trust for the organisation that holds the data. A data breach requires:

- information about a unit appearing to be learned from the data and
- information to be disseminated to others who do not already know it and are not entitled to do so.

Such information may or may not be true. When it is false the consequences for the unit and the organisation may be as serious as when it is true. Hence the importance of making it clear to the person viewing the data that it is not the original.

Techniques for ensuring that the user knows that the data being viewed is synthetic include appropriate labelling; possible approaches are detailed as the first of our checks. In addition, a synthetic data set that is much smaller than the original will make the user realise that only a subset of the data has been synthesized so that records with unique key combinations in the subsample are unlikely to identify those unique in the original.

There are two types of DR:

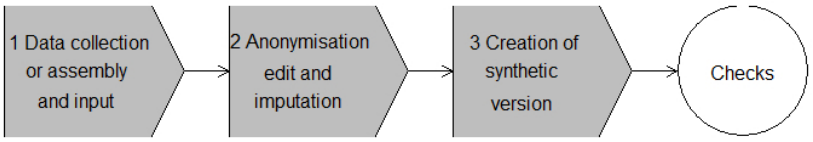
- Identity disclosure risk (IDR) – learning that an identified unit is in the data
- Attribute disclosure risk (ADR) – learning something new about a record in the original data

Each depends on what keys can be assumed to be known for subjects in the data. For people these could be (e.g. age, sex, postcode,) or for businesses (e.g. year founded, number of employees, location). The term quasi-identifiers is often used for such information. One way in which an apparent attribute disclosure can occur is when an individual is identified, so that the value of other items can be inferred from the rest of the record. Thus, by reducing IDR we will expect to reduce part of the ADR.

There may still be a real IDR for LFSD, *even when someone knows that it is synthetic*. If there is a unique value in one column of the data (e.g. exact value of a person's income or the exact date a business was founded) then someone with knowledge of that value will assume that the subject is there. These risks can be mitigated by the following steps that may often have been carried out when research-ready data is prepared (step 2 in Figure 1).

- giving data with limited precision by e.g. giving income in units of £1000 or year only for dates,

Figure 1: Steps in preparing data for synthesis



- disguising any extreme values by top or bottom coding values at the extremes of the range (e.g. outside the range of the 1st and 99th percentile), or by excluding values at the extremes of the range where the number of records outside the range is below a threshold[8],

- pooling or supressing categories with small numbers of individuals in the original (e.g. county of birth with fewer than 5 individuals).

SD may appear to pose an IDR if it includes records that have unique combinations of these keys. These records should then be checked to find if their combination of values exists in the original and is unique in the original[9]. Those unique in both the SD and the original are designated as *replicated uniques*. Where replicated uniques are a small proportion of all records they can be excluded from the SD. Nowok [23] found that the removal of *replicated uniques* from high fidelity synthetic data had very little impact on utility. If there are many records in these categories, then one or more of the keys needs to be modified to reduce this number. For example, when one of the keys is a geographic area for the unit, it can be replaced by one for a larger area (e.g. give Local Authority area rather than postcodes as the geographic key).

Our investigations show that LFSD can have a higher proportion of apparently disclosive records than the original, but only a small proportion of these are *replicated uniques*. The case study in the Appendix illustrates this (Tables A1 and A2).

Usefulness for low-fidelity synthetic data

Although the relationship between variables in the original is not maintained in the SD, there are some aspects of the SD that need to be maintained to make it useful for developing code:

- The variables in the SD must be a subset of those in the original with the same names (allowing for the addition of any prefixes or suffixes added to the SD, as described in our first check

- The levels/value labels of categorical variables must be identical except for any pooled categories in the SD.

- Whether a variable has any missing values must agree between the SD and the original.

- The precision to which any numeric data is given (e.g. number of decimal places quoted) should usually be the same in the SD and the original, unless this has been changed to reduce DR.

The metadata for the SD will usually point to the metadata for the original as the source of details, but it should also contain details of where the two differ. It should also explain which relationships in the original expect to be maintained in the SD. This will usually be

- For LFSD from codes – no tables calculated from the SD will resemble the same tables from the original.

- For LFSD from margins – only tables from the SD for one variable at a time will be similar to those from the original and the degree of similarity will depend on the detail provided about the margins. No other tables or regression model results will be expected to be similar for the SD compared to the original.

Four checks for data controllers

Here we summarise a possible protocol that might be followed to ensure LFSD will be useful for researchers and avoid any real or apparent data breaches.

1. **Label**
    a. Make it clear in the header of the metadata for the SD that this is NOT the original
    b. Ensure that the filename for the SD includes the word 'synthetic'
    c. Give variable names in the SD a prefix such as 'synth_' or a suffix such as '_synth'.
2. **Check and reduce disclosure**
    a. Identify any variables in the original where a unique value would identify an individual, such as exact values of numeric variables or rare categories for grouped variables.
    b. Define keys, as described in Section 5, and check the SD for the proportion of records in the SD that are *replicated uniques* (see above).
    c. If necessary, modify the SD with one or more of the following
        i. reduce the detail provided in the key variables to reduce the proportion of *replicated uniques* to a small proportion
        ii. reduce the precision of numeric data in the SD

[8] The second of these is more appropriate, especially for bottom coding.

[9] An alternative criterion would be to exclude records with counts of (say) 3 or fewer in the synthetic data

iii. Pool or remove rare categories for categorical data[10]

iv. remove records that are *replicated uniques*.

3. **Maintain structure**

   a. Ensure that variable names in the SD are the same as in the TRE data, allowing for 1c.

   b. Set group names for categorical data variables in the SD as exact matches to those in the TRE data, allowing for 2ciii.

   c. Check that the presence of missing values in variables agrees between original and SD.

4. **Document differences**

   a. Create metadata for the SD that points to the metadata for the TRE data and document any differences between the SD and the TRE data.

   b. Include a clear statement in the metadata describing the relationships that will expect to be maintained in the SD compared to the TRE data.

Conclusions

Our first proposed check (labelling) is probably the most important. Researchers should adhere to whatever handling instructions are provided alongside the data, such as any restrictions on sharing it. However, careful labelling of SD by the creator can mitigate the ensuing risks, even if handling instructions are not followed and wider sharing does occur. Labelling is even more critical where SD are freely available without the need to register and agree to sharing restrictions.

While LFSD will be expected to have a low DR, this risk may not be zero. For example, metadata may give the range of actual values for numeric variables and the highest of lowest value may identify a small number of individuals. A similar situation can occur for an infrequent code for a categorical variable. Our proposals on checking for *replicated uniques* are based on recent work for high fidelity SD [13]. Some people may argue that they are not necessary for LFSD, but we think that some attempt should be made to measure these risks, especially for data that is relatively freely available. A possible exception might be LFSD created from codes where DR is low both a-priori and from our experience, as exemplified in the case study.

Steps 3 and 4 are important to make the LFSD useful to researchers. There are other aspects of usefulness that we have not specified as a necessary check, but can be very beneficial to researchers. An example would be providing resources to help users create code to add labels to categorical variables.[11]

We hope that the proposed checks for data holders who plan to supply LFSD to potential users will be useful and will play a part in allowing administrative data to provide information for public benefit.

[10] A referee has pointed out that the absence of a code might itself be a disclosure risk, suggesting that such procedures might be better carried out on the data before synthesis.

[11] The provision of JSON files is one option, see https://en.wikipedia.org/wiki/JSON.

Acknowledgments

We would like to thank two anonymous referees whose insightful comments helped us to improve this paper. Part of Gillian Raab's time in 2023/24 to develop DR measures for synthetic data was supported by Research Data Scotland.

Statement on conflicts of interest

None that we are aware of.

Ethics statement

Although some of the individual studies that have informed our practice have required ethical permission, none was required for this paper because no data from individuals or organisations was used in its preparation. We hope that readers will think that the checks we propose in this paper will go some way to answer ethical concerns for LFSD.

Data availability statement

The data analysed in the Appendix are derived from historic Census data made available by the Integrated Census Microdata Project. This data is freely available to download from https://www.campop.geog.cam.ac.uk/research/projects/icem/ as it no longer pertains to any living individuals. The extract used to illustrate methods in the Appendix can also be downloaded from www.gillianraab.co.uk/1901census, along with the code used to prepare the data and carry out the analyses described in the Appendix.

References

1. Sudlow C (2024) *Uniting the UK's Health Data: A Huge Opportunity for Society*. HDR UK, web https://www.hdruk.ac.uk/helping-with-health-data/the-sudlow-review/, accessed 4/3/2025. https://doi.org/10.5281/zenodo.13353746

2. Penner AM, Dodge KA. (2019) *Using administrative data for social science and policy*. J Soc Sci. 2019;5(3):1–18. https://doi.org/10.7758/RSF.2019.5.3.01, accessed 4/3/2025.

3. Nowok B., Raab G.M., Dibben C, (2017) *Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R*. Statistical Journal of the IAOS, 33(3):785–796; https://doi.org/10.3233/SJI-150153, accessed 4/3/2025.

4. Dibben, C, Raab G.M., Nowok, B., Williamson, L., Adair, L. (2024) *Synthpop: A Tool to Enable More Flexible Use of Sensitive Data within the Scottish Longitudinal Study* in Drechsler, Jörg.(ed) Handbook of Sharing Confidential Data: Differential Privacy, Secure Multiparty Computation, and Synthetic Data. S.l: CHAPMAN and HALL CRC, 2024.4. https://doi.org/10.1201/9781003185284

5. Van Kesteren, E (2024) *To democratize research with sensitive data, we should make synthetic data more accessible*, Patterns, Volume 5, Issue 9. https://doi.org/10.1016/j.patter.2024.101049, accessed 4/3/2025.

6. Rubin DB. (1993) *Statistical disclosure limitation*. J Off Stat. 1993;9(2):461–8.

7. Reiter, J.: *Synthetic data: A look back and a look forward*. Transactions in Data Privacy 16, 15—–24

8. Drechsler and Haensch (2024) Drechsler, J., Haensch, A.C.: 30 years of synthetic data. Statist. Sci. 39(2), 221–242. https://doi.org/10.48550/arXiv.2304.02107

9. Bulmer, M., & Coote, L. (2022). The role of synthetic data in teaching and learning statistics. In *Bridging the gap: empowering & educating today's learners in statistics*. Proceedings of the 11th international conference on teaching statistics. https://iase-web.org/icots/11/proceedings/pdfs/ICOTS11_422_BULMER.pdf, accessed 4/3/2025. https://doi.org/10.52041/iase.icots11.T14I2

10. Kokosi T, De Stavola B, Mitra R, Frayling L, Doherty A, Dove I, Sonnenberg P, Harron K. *An overview of synthetic administrative data for research*. Int J Popul Data Sci. 2022 May 23;7(1):1727. https://ijpds.org/article/view/1727, accessed 4/3/2025. https://doi.org/10.23889/ijpds.v7i1.1727

11. Office for National Statistics (2021). *ONS methodology working paper series number 16 - Synthetic data pilot*. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot, accessed 4/3/2025.

12. Little, C., Elliot, M., and Allmendinger, R. (2022) *Comparing the utility and disclosure risk of synthetic data with samples of microdata*. In Privacy in Statistical Databases 2022 J. Domingo-Ferrer and M. Laurent, Eds., Springer International Publishing, pp. 234–249. https://doi.org/10.48550/arXiv.2207.03339

13. Raab, GM, Nowok B, and Dibben C. (2024), *Practical Privacy Metrics for Synthetic Data*. Available as a vignette in versions of the synthpop package. form 1.9-0 onwards. https://cran.r-project.org/web/packages/synthpop/vignettes/disclosure.pdf, accessed 11/6/2025. https://doi.org/10.48550/arXiv.2406.16826

14. Calcraft P, Thomas I, Maglicic M, Sutherland A.(2021) ADRUK, *Accelerating public policy research with synthetic data*. Available from: https://www.adruk.org/fileadmin/uploads/adruk/Documents/Accelerating_public_policy_research_with_synthetic_data_December_2021.pdf.

15. ADR UK (2023) An interim ADR UK position statement on synthetic data https://www.adruk.org/fileadmin/uploads/adruk/Documents/An_interim_ADR_UK_position_statement_on_synthetic_data.pdf, accessed 4/3/2025.

16. Research Data Scotland (2024), Synthetic census data now available for research. https://www.researchdata.scot/news-and-insights/synthetic-census-data-now-available-for-research/, accessed 22/12/024.

17. NHS England, Artificial Data Pilot, https://digital.nhs.uk/services/artificial-data#random-data-generation, accessed 11/6/2025.

18. Bharat SS, Frayling L, Stock J, Lugg-Widge F, Gordon E, Oliver E. A Review of Synthetic Data Terminology for Privacy Preserving Use Cases, Int J Popul Data Sci. 2025 (submitted).

19. Grath-Lone LM, Jay MA, Blackburn R, Gordon E, Zylbersztejn A, Wiljaars L, Gilbert R. *What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature*. Int J Popul Data Sci. 2022 Apr 27;7(1):1718. https://ijpds.org/article/view/1718, accessed 4/3/2025. https://doi.org/10.23889/ijpds.v7i1.1718

20. Elliot, M., Mackey, E., and O'Hara, K. (2020) The anonymisation decision-making framework: European practitioners. https://ukanon.net/framework/, accessed 2022-02-23.

21. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P. P. (2012). *Statistical disclosure control*. John Wiley & Sons. https://doi.org/10.1111/anzs.12085

22. Templ, M (2017). *Disclosure Control for Microdata: Methods and Applications in R*. 1st ed. 2017. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50272-4

23. Nowok B, Raab GM, Dibben C. (2017) *Recognising real people in synthetic microdata: risk mitigation and impact on utility*. Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality; Skopje, North Macedonia, 20-22 September 2017. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/3_risk_mitigation.pdf, accessed 11/6/25.

24. Taub J, Elliot M, Raab GM, Chareset A, Chen C, O'Keefe CM, Pistner M, Snoke J, Slavkovic A (2019) *Creating the Best Risk-Utility Profile: The Synthetic Data Challenge* , Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf, accessed 11/6/25.

Appendix: case study from 1901 Census of Scotland

To illustrate the creation of LFSD we need access to the original data. Here we use an extract from the 1901 Census of Scotland, now free from any disclosure issues. The source of the data is the Integrated Census Microdata Project[12]. The records used in our analysis are a subset of variables used for the Newton Institute synthetic data challenge in 2017 [24]. The data, metadata and the code used for the analyses presented in this paper are available[13]. We used 82,851 records for heads of households in Midlothian (the area that includes the City of Edinburgh).

Possible keys to use as quasi-identifiers included an area level (parish) with 29 levels that can be grouped into 3 larger areas. We investigated identity disclosure for these data based on 4 choices of keys as shown below,

Keys 1	sex, grouped parish ,disability	No uniques: 4 (0.005%)
Keys 2	sex, parish ,disability	No uniques: 33 (0.04%)
Keys 3	sex, parish ,disability, occlab1[14]	No uniques: 241 (0.3%)
Keys 4	sex, parish ,disability, occlab1, age	No uniques: 7096 (8.6%)

The keys were chosen as plausible quasi-identifiers to provide a range of identity DR for the original data. Tables A1 and A2 show results for the full data set and for a subset restricted to employers[15]. For these examples the synthesis from codes and ranges reduces the level of *replicated uniques* to zero for the first two keys, and to below 1% for the other two. Synthesis from exact marginals seems to be effective at reducing *replicated uniques* for the first two sets of keys, but less so for the other two, especially the last one. If these keys were thought to be appropriate quasi-identifiers, then some action to reduce the disclosure risk from the original data should be carried out[16].

Further analyses were carried out to investigate the identity *d*isclosure risk from releasing a synthetic data set with fewer records than the original data. The full 1901 Census data of *size n = 5,364* was used to create synthetic samples of size *n/k*,[17] where *k* can take on values 1,2,3, ... 10. Results are shown in Figure A1.

Synthetic data has higher proportions of unique records on account of its smaller sample size, but this is not true for *replicated uniques* from the synthetic data, where smaller sizes of the synthetic data have either a similar rate of *replicated uniques* or, in some cases a lower rate.

We cannot generalise from this single case study. Further work is indicated to investigate the formal disclosure risk for SD, according to the properties of the original data and the methods used to create the SD.

Table A1: Numbers and percentages of unique records and replicated unique records for the full 1901 Census data (n = 82,851). Average results for 500 syntheses

	Original		**Synthesised from only codes and ranges**				**Synthesised from exact marginal distributions**			
			Synthetic		**Replicated uniques**		**Synthetic**		**Replicated uniques**	
	n	**% unique**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
keys1	4	0.005	0	0	0	0	5.0	0.006	0.8	0.001
keys2	33	0.040	0	0	0	0	37.3	0.045	7.4	0.009
key 3	241	0.291	16.6	0.02	0.4	0.001	289.0	0.349	35.6	0.043
keys4	7096	8.565	75383.2	91.00	609.6	0.736	10148.0	12.25	964.7	1.164

[12]See https://www.campop.geog.cam.ac.uk/research/projects/icem/, accessed 9/6/2025.

[13]See www.gillianraab.co.uk/1901census.

[14]The highest grouping of the historic classification of occupations.

[15]Note that this approach is necessary because disclosure risk refers to the ability to identify a unit in the original data. A random sample of the original would not be appropriate because possible units in the original data would not be known to an intruder.

[16]For this example, grouping the values of age would be the most obvious step.

[17]Rounded to the nearest whole number.

Table A2: Numbers and percentages of unique records and replicated unique records for the 1901 Census data, restricted to employers (n = 5,364). Average results for 500 syntheses

	Original		Synthesised from only codes and ranges				Synthesised from exact marginal distributions			
			Synthetic		Replicated uniques		Synthetic		Replicated uniques	
	n	% unique	n	%	n	%	n	%	n	%
keys1	1	0.019	0	0	0	0	0.6	0.012	0.3	0.006
keys2	6	0.112	0	0	0	0	6	0.112	2	0.037
keys3	152	2.834	781.4	14.6	42.8	0.798	187	3.487	36.6	0.683
keys4	1699	31.67	5211.6	97.2	47.5	0.886	1739.7	32.43	269.1	5.017

Figure A1: Percentage uniques in synthetic records and percentage replicated uniques by number of records in the synthetic data. Results are averaged over 500 syntheses. The solid black lines are the percentage of unique key combinations in the original data.