Generating Combined Synthetic Data from Distributed Analytical Outputs

Claire Little, Richard Allmendinger and Mark Elliot

Abstract

This paper examines the feasibility of generating coherent, combined synthetic datasets (for teaching, evaluation, and exploratory research) using only cleared analytical outputs from multiple independent data owners. Using an evolutionary algorithm (EA), analytical outputs such as regression coefficients and summary tables are used to guide the synthesis of data without direct access to any safeguarded microdata. A series of experiments explore scenarios with varying degrees of overlap in variables and samples, scaling from two to four datasets. Across all cases, the resulting synthetic datasets closely reproduce original analytical outputs and consistently outperform approaches that merge independently synthesised datasets. These findings demonstrate the robustness and scalability of a federated-style approach to synthetic data.

1. Introduction

Access to detailed microdata is increasingly constrained by legal, ethical, and governance requirements designed to protect confidentiality and prevent disclosure. While such controls are essential, they present challenges for teaching, methodological development, and exploratory research, where flexible access to realistic data structures is often required. Synthetic data has emerged as a promising solution, offering data that resemble real datasets in structure and statistical properties while avoiding direct disclosure of sensitive information.

Most synthetic data generation approaches assume full access to the dataset during the synthesis process (see [1] for a review). However, in many real-world settings—particularly those involving multiple data owners or secure research environments—such access is not feasible. Instead, data custodians may only be able to release cleared analytical outputs, such as regression coefficients or summary tables. This raises a critical question: can analytically useful synthetic datasets be generated using only these limited outputs, without any access to the original microdata?

This paper addresses that question by investigating a federated-style approach to synthetic data generation, in which analytical outputs from multiple independent datasets are combined to guide the synthesis of a single, coherent synthetic dataset. Using an evolutionary algorithm (EA) [2], the approach seeks to reproduce released analytical results while preserving consistency across outputs derived from different data sources. The work is situated within the context of teaching and methodological use cases, where analytical utility and interpretability are often more important than exact record-level fidelity.

Through a series of experiments that vary the number of contributing datasets, the degree of overlap in variables, and the overlap in samples, this study evaluates the feasibility, robustness, and scalability of the proposed approach. Particular attention is paid to whether joint synthesis from combined outputs outperforms approaches that independently synthesise datasets (and subsequently attempt to merge them).

2. Literature Review

2.1 Synthetic Data Generation

Synthetic data generation has a long history in statistical disclosure control [1], where it has been used to reduce disclosure risk [3] while retaining analytical utility [4]. Early approaches focused on parametric modelling and multiple imputation frameworks [5,6], generating synthetic records from fitted statistical models [7]. More recent developments have incorporated machine learning techniques, including decision trees [8], Bayesian networks [9], and deep generative models such as Generative Adversarial Networks (GANs) [10,11].

While these methods can produce realistic data, they typically require full access to the original dataset during training. This assumption limits their applicability in settings where data cannot be pooled or where only aggregate information can be released. Moreover, high-fidelity generative models may introduce new disclosure risks if not carefully controlled, particularly when overfitting occurs [12].

2.2 Utility-Driven and Output-Constrained Synthesis

An alternative strand of research emphasises utility-driven synthesis, where synthetic data are explicitly optimised to reproduce specific analytical results rather than the full joint distribution of the data [13, 14]. In this framework, analytical outputs—such as regression coefficients or contingency tables—serve as constraints or targets for the synthesis process. EAs and other optimisation-based methods have been shown to be effective in this context, particularly when the desired outputs are well defined.

Output-constrained synthesis aligns closely with the needs of Trusted Research Environments (TREs), where cleared outputs are routinely produced having been scrutinised for disclosure risk. By limiting the information used in synthesis to already approved outputs, this approach offers a transparent and governance-friendly pathway to synthetic data creation. However, our existing work considers only single datasets and does not address the challenges of combining outputs from multiple sources.

2.3 Federated and Distributed Data Contexts

Federated analysis [15] and learning [16, 17] have gained prominence as mechanisms for extracting value from distributed data without centralising sensitive information. In such settings, models or summary statistics are shared rather than raw data. While federated learning focuses on training shared predictive models, less attention has been paid to the generation of shared synthetic datasets from federated outputs. Although our recent review [18] indicates growing interest.

The generation of combined synthetic data from multiple independent outputs introduces additional challenges, including ensuring coherence across analyses, resolving differences in sample composition, and managing partial overlap in variables. Naïvely combining synthetic datasets generated independently from each source does not guarantee analytical consistency, particularly when relationships span datasets.

2.4 Contribution of This Paper

This paper contributes to the literature by demonstrating that coherent synthetic datasets can be generated directly from combined analytical outputs released by multiple data owners. By

using an EA to jointly satisfy multiple analytical constraints, the approach avoids the pitfalls of post hoc dataset merging and scales to scenarios involving increasing numbers of datasets.

The focus on teaching and methodological use cases positions this work as a practical complement to existing synthetic data research, offering a low-risk, transparent, and governance-compatible method for producing analytically faithful synthetic data in restricted access environments.

3. Methodological Framework

This section describes the methodological framework used to generate a single coherent synthetic dataset from analytical outputs released by multiple independent data owners. The framework is designed for settings in which direct access to microdata is not possible and only cleared analytical outputs and limited metadata are available. Released outputs are treated as explicit utility constraints, and an EA is used to jointly satisfy these constraints during data synthesis.

3.1 Problem Setting and Assumptions

We consider a setting in which multiple data owners each hold a dataset that cannot be shared or pooled due to governance, confidentiality, or legal restrictions. Let there be K independent data owners, each holding a dataset

$$D_k = \{\mathbf{x}_{k1}, \mathbf{x}_{k2}, \dots, \mathbf{x}_{kn_k}\}, k = 1, \dots, K,$$

where each $\mathbf{x}_{ki}$ denotes an individual record represented as a vector of variable values, and the microdata D_k are not accessible outside the owner's secure environment.

Each data owner releases a set of cleared analytical outputs

$$\mathcal{O}_k = \{o_{k1}, \dots, o_{km_k}\},$$

derived from known analytical procedures applied to D_k. These outputs may be based on overlapping or non-overlapping variable sets, and the underlying datasets may differ in sample composition.

The objective is to generate a single synthetic dataset

$$\widetilde{D} = \{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{\tilde{n}}\},$$

such that applying the same analytical procedures to $\widetilde{D}$ yields outputs that closely reproduce all released outputs $\mathcal{O}_1, \dots, \mathcal{O}_K$. The synthetic dataset is not intended to approximate the full joint distribution of any D_k, but rather to reproduce released analytical results to an acceptable level of error.

Accordingly, the framework operates under the assumptions that: (i) no access to original microdata is available at any stage; (ii) only cleared analytical outputs and non-disclosive metadata are used; and (iii) the resulting synthetic dataset must be internally coherent and analytically interpretable.

3.2 Analytical Outputs as Utility Constraints

Each released analytical output defines a set of utility constraints on the synthetic data. Applying the analytical function $f_{kj}(\cdot)$ to the synthetic dataset yields

$$\tilde{o}_{kj} = f_{kj}(\tilde{D}).$$

The synthesis task is therefore framed as a joint optimisation problem in which the synthetic dataset is adjusted so that $\tilde{o}_{kj} \approx o_{kj}$ for all outputs across all data owners. This utility-driven formulation avoids attempting to reconstruct the full data-generating process and instead focuses explicitly on reproducing the analyses that users are expected to perform.

When multiple outputs are released—potentially from different datasets—the framework seeks a single synthetic dataset that jointly satisfies all constraints. This distinguishes the approach from methods that independently synthesise datasets and attempt to merge them post hoc.

3.3 Metadata and Variable Representation

In addition to analytical outputs, limited non-disclosive metadata are assumed to be available. Let $\mathcal{M}_k$ denote metadata for dataset D_k, including variable definitions, category labels, and marginal proportions.

To ensure a consistent representation, combined metadata $\mathcal{M}^*$ are constructed. For categorical variables, category proportions are averaged across datasets:

$$p_c^* = \frac{1}{K}\sum_{k=1}^{K} p_{kc} \quad ,$$

where p_{kc} is the proportion of category c in dataset D_k.

The combined metadata constrain the feasible space of synthetic datasets and ensure realistic structure, while introducing no additional disclosure risk.

3.4 Evolutionary Algorithm for Joint Synthesis

An EA is used to search for a synthetic dataset that satisfies all analytical constraints. Each candidate solution represents a complete synthetic dataset $\tilde{D}^{(g,i)}$, where g indexes generations and i indexes individuals within a generation.

For each candidate dataset, a fitness function is defined as:

$$\mathcal{L}(\tilde{D}) = \sum_{k=1}^{K} \sum_{j=1}^{m_k} w_{kj}\, \ell\,(o_{kj}, \tilde{o}_{kj}),$$

where $\ell(\cdot,\cdot)$ is the loss function and w_{kj} are optional weights.

The EA proceeds via initialisation from $\mathcal{M}^*$, evaluation, selection, crossover, and mutation. A stepped mutation schedule is used to encourage early exploration and later convergence. The

stepped mutation schedule divides the mutation rate by three at regular intervals and in preliminary experiments was more effective than a fixed mutation rate. This optimisation-based approach is well suited to handling multiple, potentially competing analytical constraints.

3.5 Evaluation Metrics

During optimisation, mean squared error (MSE) is used as the fitness measure:

$$\text{MSE} = \frac{1}{|\mathcal{O}|}\sum_{k,j}(o_{kj} - \tilde{o}_{kj})^2 \ .$$

For reporting and interpretation, mean absolute error (MAE) is used:

$$\text{MAE} = \frac{1}{|\mathcal{O}|}\sum_{k,j}|\ o_{kj} - \tilde{o}_{kj}\ |\ .$$

Together, these metrics provide a clear assessment of optimisation performance and analytical fidelity across experiments.

4. Data and Experimental Design

This section describes the data used in the study and the experimental design employed to evaluate the proposed joint synthesis framework. The experiments are designed to assess whether analytically coherent synthetic datasets can be generated from distributed analytical outputs, and how performance varies with the number of contributing datasets and the degree of overlap in samples and variables.

4.1 Data Description

The experiments are based on a set of related datasets sharing a common variable structure but differing in sample composition. All variables are categorical, reflecting typical data released for teaching and methodological purposes in secure environments. The outcome variable is binary, and covariates include a set of categorical predictors commonly used in applied regression analysis.

The original datasets are not accessed directly during synthesis. Instead, all experiments rely exclusively on cleared analytical outputs and non-disclosive metadata derived from these datasets. Metadata specify variable definitions, category labels, and marginal category proportions. No record-level information or joint distributions are used at any stage.

To support joint synthesis, combined metadata are constructed as described in Section 3.3, ensuring that all candidate synthetic datasets share a common variable representation.

For simplicity we are using one dataset (a subset of the 1991 SARS UK Census dataset [19]) and partitioning that up to represent the different datasets.

The overall dataset, the 1991 SARS UK Census dataset was subsetted on seven areas in the West Midlands region (using the AREAP variable) and used eight variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, QUALNUM, SEX, TENURE). TENURE (housing tenure)

was used as the target variable for regression models, and this was converted to binary (individual lives in an owned or rented house). There were 50357 records in the dataset, after removing those with missing values for TENURE. Appendix A1 contains a data dictionary, and Appendix A2 details the reference levels for the regression models.

4.2 Experimental Scenarios

A sequence of experimental scenarios is constructed to examine how synthesis performance changes as the number and structure of contributing datasets vary. The scenarios are ordered to progressively increase complexity.

- Experiment 1: Two datasets with different samples/rows and the same variables
- Experiment 2: Three datasets with different samples/rows and the same variables
- Experiment 3: Four datasets with different samples/rows and the same variables
- Experiment 4: The data is first split into an original and holdout dataset, then from the original two datasets are created each with the same samples/rows but different variables (with the degree of overlap between outputs tested)

Each scenario is evaluated using the same synthesis framework and evaluation metrics, allowing direct comparison across experiments.

4.3 Synthetic Data Generation Procedure

For all experiments the EA was initialised using a population (size = 32) of synthetic datasets (candidates) derived from the uniform distribution of the original (training) data. The fitness of these datasets (candidates) was measured by performing the same logistic regression (using the same data pre-processing steps as performed on the original data) and comparing the resulting regression coefficients to the original coefficients. The fitness was calculated at each generation. Fitness was calculated as the MSE between synthetic and original coefficients. The fitness score should tend towards zero, i.e., a lower fitness score is optimal (a score of zero would indicate that the logistic regression model had been recreated identically).

Exploratory experiments determined that crossover with stepped mutation produced the best solution (stepped mutation is where the mutation rate is divided by three every 250 generations), and the EA ran for 2000 generations. Ten randomly initialised runs of the EA were performed. At each generation the MAE between the regression coefficients (for the optimal dataset) and the original was recorded.

4.4 Analytical Outputs Used

One class of analytical outputs are used to define utility constraints:

- **Regression outputs**: Coefficients from logistic regression models fitted to each dataset, using a common model specification where applicable.

This reflects analyses that are routinely permitted for release from secure research environments and are sufficient to capture both multivariate and marginal relationships relevant for teaching use cases.

Results are summarised using regression coefficient comparisons, tabular discrepancies, and graphical summaries. All figures and tables are interpreted in terms of analytical fidelity rather than record-level similarity, consistent with the intended use of the synthetic data.

5. Experimental Results

This section uses a mix of visuals and tables to present and analyse our results including: (i) EA convergence plots (Figure 1) and (ii) regression coefficient tables (e.g. Tables 4 and 5) to demonstrate stable optimization behaviour across all experiments consistently low mean absolute error (MAE) between original and synthetic outputs, respectively, (iii) comparative tables (e.g. Table 6) to highlight that direct synthesis from combined analytical outputs yields synthetic datasets closest to the original full dataset, and to further demonstrate that descriptive structure is well preserved, even when outputs are distributed across datasets with limited overlap.

5.1 Experiment 1

This scenario has two datasets each with different samples/rows and the same variables. For this we split our original dataset into two (on the rows), with dataset A having 25178 rows and dataset B having 25179 rows. Each have the same eight variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, QUALNUM, SEX, TENURE).

The analytical output for was a logistic regression performed on the TENURE variable, with all remaining variables used as predictors. The same logistic regression was performed on datasets A and B; it was also performed on the original dataset for comparison. Figure 1 plots the MSE fitness for each of the ten randomly initialised runs of the EA.

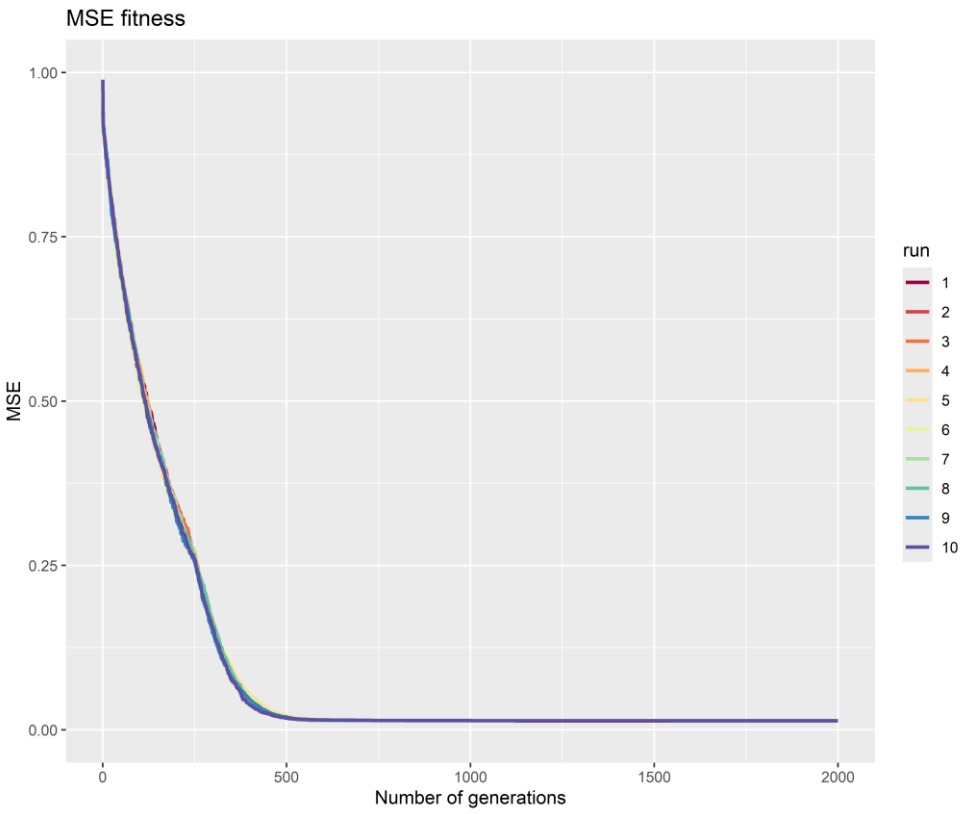

Figure 1: Experiment 1, MSE fitness plot for ten randomly initialised runs

To compare the results, we consider the error for the original datasets A and B compared to the original, and then consider the synthetic data compared to the original. In Table 1**Error! Reference source not found.** we can see that the model trained on the synthetic data has lower error than each of the models trained on the individual original datasets.

Table 1: Experiment 1, regression coefficients for synthetic data and original datasets A and B, compared to the original regression coefficients

	Dataset A and B combined)	Original Dataset A	Original Dataset B	Synthetic
Features	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**
Constant	0.258	0.315	0.203	0.256
AGE	0.002	0.001	0.003	0.006
AREAP_58	-0.386	-0.411	-0.361	-0.385
AREAP_59	-0.364	-0.399	-0.329	-0.362
AREAP_60	0.299	0.309	0.290	0.299
AREAP_61	-0.733	-0.762	-0.705	-0.732
AREAP_62	0.053	0.021	0.083	0.054
AREAP_63	0.134	0.120	0.147	0.134
ETHGROUP_10	0.780	0.703	0.859	0.785
ETHGROUP_2	0.490	0.373	0.600	0.493
ETHGROUP_3	1.158	1.114	1.217	1.168
ETHGROUP_4	0.914	1.174	0.641	0.894
ETHGROUP_5	-1.658	-1.669	-1.648	-1.658
ETHGROUP_6	-0.829	-0.846	-0.812	-0.828
ETHGROUP_7	-0.029	-0.159	0.100	-0.023
ETHGROUP_8	0.640	0.795	0.485	0.631
ETHGROUP_9	0.369	0.283	0.446	0.369
LTILL_2	-0.646	-0.680	-0.614	-0.645
MSTATUS_2	-0.824	-0.731	-0.914	-0.827
MSTATUS_3	-0.381	-0.351	-0.411	-0.383
MSTATUS_4	0.396	0.482	0.312	0.393
MSTATUS_5	-0.066	-0.028	-0.103	-0.067
QUALNUM_1	-0.983	-0.939	-1.024	-0.984
QUALNUM_2	-1.523	-1.667	-1.387	-1.520
SEX_2	0.056	0.039	0.074	0.057
MAE between dataset and whole original:		0.0634	0.0629	0.0032

5.2 Experiment 2

This scenario has three datasets each with different samples/rows and the same variables. For this we split our original dataset into three (on the rows), with dataset A and B having 16786 rows and dataset C having 16785 rows. Each have the same eight variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, QUALNUM, SEX, TENURE).

The analytical output for was a logistic regression performed on the TENURE variable, with all remaining variables used as predictors. The same logistic regression was performed on datasets A, B and C, it was also performed on the original dataset for comparison.

To compare the results, we consider the error for the original datasets A, B and C compared to the original, and then consider the synthetic data compared to the original. In Table 2**Error! Reference source not found.** we can see that the model trained on the synthetic data has lower error than each of the models trained on the individual original datasets.

Table 2: Experiment 2, regression coefficients for synthetic data and original datasets A, B and C, compared to the original regression coefficients

	Original (Dataset A, B, C combined)	Original Dataset A	Original Dataset B	Original Dataset C	Synthetic
Features	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**
Constant	0.258	0.191	0.265	0.322	0.262
AGE	0.002	0.002	0.001	0.003	0.006
AREAP_58	-0.386	-0.412	-0.414	-0.338	-0.387
AREAP_59	-0.364	-0.379	-0.377	-0.337	-0.364
AREAP_60	0.299	0.316	0.265	0.319	0.299
AREAP_61	-0.733	-0.788	-0.684	-0.732	-0.732
AREAP_62	0.053	0.038	0.084	0.039	0.055
AREAP_63	0.134	0.184	0.136	0.085	0.134
ETHGROUP_10	0.780	0.742	1.059	0.590	0.802
ETHGROUP_2	0.490	0.561	0.507	0.395	0.485
ETHGROUP_3	1.158	1.474	0.592	1.312	1.107
ETHGROUP_4	0.914	1.178	0.758	0.831	0.911
ETHGROUP_5	-1.658	-1.661	-1.730	-1.591	-1.662
ETHGROUP_6	-0.829	-0.972	-0.754	-0.762	-0.824
ETHGROUP_7	-0.029	0.027	-0.046	-0.071	-0.032
ETHGROUP_8	0.640	0.550	0.661	0.708	0.643
ETHGROUP_9	0.369	0.293	0.308	0.543	0.383
LTILL_2	-0.646	-0.618	-0.628	-0.696	-0.648
MSTATUS_2	-0.824	-0.794	-0.786	-0.892	-0.824
MSTATUS_3	-0.381	-0.344	-0.427	-0.378	-0.385
MSTATUS_4	0.396	0.385	0.393	0.412	0.397
MSTATUS_5	-0.066	-0.075	0.006	-0.125	-0.063
QUALNUM_1	-0.983	-0.927	-1.086	-0.937	-0.986
QUALNUM_2	-1.523	-1.508	-1.549	-1.531	-1.530
SEX_2	0.056	0.114	0.031	0.024	0.054
MAE between outputs and whole original:		0.0618	0.0704	0.0578	0.0058

5.3 Experiment 3

This scenario has four datasets each with different samples/rows and the same variables. For this we split our original dataset into four (on the rows), with dataset A 12590 rows and

datasets B, C and D having 12589 rows. Each have the same eight variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, QUALNUM, SEX, TENURE).

The analytical output for this was a logistic regression performed on the TENURE variable, with all remaining variables used as predictors. The same logistic regression was performed on datasets A, B, C and D, it was also performed on the original dataset for comparison.

To compare the results, we consider the error for the original datasets A, B, C and D compared to the original, and then consider the synthetic data compared to the original. In Table 3**Error! Reference source not found.** we can see that the model trained on the synthetic data has lower error than each of the models trained on the individual original datasets.

Table 3: Experiment 3, regression coefficients for synthetic data and original datasets A, B, C and D, compared to the original regression coefficients

	Coefficients					
Features	Dataset A, B, C, D combined	Original Dataset A	Original Dataset B	Original Dataset C	Original Dataset C	Synthetic
Constant	0.258	0.217	0.212	0.354	0.250	0.262
AGE	0.002	0.005	0.003	0.000	0.001	0.007
AREAP_58	-0.386	-0.504	-0.434	-0.321	-0.281	-0.390
AREAP_59	-0.364	-0.388	-0.340	-0.363	-0.363	-0.365
AREAP_60	0.299	0.242	0.270	0.388	0.296	0.300
AREAP_61	-0.733	-0.662	-0.789	-0.750	-0.739	-0.731
AREAP_62	0.053	-0.020	-0.001	0.168	0.063	0.053
AREAP_63	0.134	0.170	0.133	0.117	0.115	0.136
ETHGROUP_10	0.780	0.428	0.720	0.842	1.103	0.750
ETHGROUP_2	0.490	0.441	0.363	0.621	0.541	0.496
ETHGROUP_3	1.158	0.916	1.808	0.333	1.584	1.090
ETHGROUP_4	0.914	0.479	1.154	0.881	1.167	0.885
ETHGROUP_5	-1.658	-1.452	-1.666	-1.757	-1.790	-1.656
ETHGROUP_6	-0.829	-0.833	-0.827	-0.847	-0.809	-0.831
ETHGROUP_7	-0.029	-0.113	0.169	-0.451	0.208	-0.078
ETHGROUP_8	0.640	1.150	0.416	0.349	0.496	0.627
ETHGROUP_9	0.369	0.461	0.074	0.389	0.660	0.399
LTILL_2	-0.646	-0.598	-0.641	-0.719	-0.626	-0.646
MSTATUS_2	-0.824	-0.932	-0.824	-0.740	-0.804	-0.827
MSTATUS_3	-0.381	-0.502	-0.367	-0.423	-0.245	-0.396
MSTATUS_4	0.396	0.292	0.507	0.516	0.273	0.398
MSTATUS_5	-0.066	-0.274	-0.113	0.070	0.059	-0.072
QUALNUM_1	-0.983	-0.885	-1.186	-0.918	-0.937	-0.971
QUALNUM_2	-1.523	-1.495	-1.489	-1.491	-1.626	-1.520
SEX_2	0.056	0.032	0.091	0.055	0.048	0.055
MAE between outputs and whole original:		0.1255	0.1005	0.1143	0.1045	0.0116

5.4 Experiment 4

This is a variation where the data is first split into two datasets – a train (or original) and holdout. This was a fifty-fifty split, with train having 25178 records and the holdout having 25179. The holdout is set aside and not used during the EA process but used to provide a holdout comparison at the end. The training dataset is then split into two datasets (each with the same samples/rows but different variables, and at least one variable overlapping both). Experiment 4.1 has one overlapping variable, 4.2 has two overlapping variables and 4.3 has three overlapping variables. The holdout will be used to evaluate the synthetic data by calculating how a logistic regression created from the synthetic data predicts on the unseen holdout data, compared to the logistic regressions created from datasets A and B separately.

5.4.1 Experiment 4.1

Overall, there were 8 variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, QUALNUM, SEX, TENURE) – the datasets (n=25178) were split such that

- Dataset A had AREAP, ETHGROUP, QUALNUM, SEX, TENURE
- Dataset B had AGE, LTILL, MSTATUS, TENURE

TENURE was therefore the overlapping variable. For both dataset A and B a logistic regression was the analytical output, with TENURE being used as the target variable and the remaining variables as predictors. Run 2 produced the optimal dataset, with the lowest fitness. Table 4 compares the logistic regression coefficients for output A to those obtained from the synthetic data produced by run 2. Table 5 compares the logistic regression coefficients for output B to those obtained from the synthetic data produced by run 2.

Table 4: Experiment 4.1, regression coefficients for synthetic run 2 compared to the original regression A coefficients

Feature	Original Coefficient	Synthetic Coefficient	Absolute difference
Constant	-0.499	-0.398	0.101
AREAP_58	-0.331	-0.318	0.012
AREAP_59	-0.490	-0.479	0.011
AREAP_60	0.250	0.264	0.014
AREAP_61	-0.793	-0.786	0.008
AREAP_62	0.008	0.016	0.007
AREAP_63	0.109	0.120	0.010
ETHGROUP_10	0.858	0.860	0.003
ETHGROUP_2	0.648	0.654	0.006
ETHGROUP_3	0.769	0.775	0.006
ETHGROUP_4	1.271	1.276	0.005
ETHGROUP_5	-1.646	-1.644	0.002
ETHGROUP_6	-0.878	-0.871	0.007
ETHGROUP_7	-0.124	-0.116	0.008

ETHGROUP_8	0.634	0.637	0.004
ETHGROUP_9	0.491	0.494	0.003
QUALNUM_1	-1.085	-1.071	0.015
QUALNUM_2	-1.621	-1.611	0.010
SEX_2	0.043	0.094	0.051
		MAE:	**0.015**

Table 5: Experiment 4.1, regression coefficients for synthetic run 2 compared to the original regression B coefficients

Feature	Original Coefficient	Synthetic Coefficient	Absolute difference
Constant	0.112	0.112	0.000
AGE	0.005	0.001	0.004
LTILL_2	-0.707	-0.709	0.001
MSTATUS_2	-0.954	-0.953	0.001
MSTATUS_3	-0.464	-0.463	0.001
MSTATUS_4	0.320	0.320	0.000
MSTATUS_5	-0.164	-0.164	0.000
		MAE:	**0.001**

The holdout data was put through the logistic regression models from output A and output B to measure their accuracy in predicting TENURE. These models were generated using the original data. A logistic regression model was then generated using the synthetic data and the holdout data put through this. The accuracy for all models is in Table 6. The model built on the synthetic dataset outperforms models A and B, both of which were trained on the original. The synthetic data was therefore able to predict better on a holdout dataset than models trained on the two original datasets.

Table 6: Experiment 4.1, logistic regression model accuracy on holdout data

Model	Accuracy
Output A	0.672
Output B	0.678
Synthetic	0.686

5.4.2 Experiment 4.2

Overall, there were 8 variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, QUALNUM, SEX, TENURE) – the datasets (n=25178) were split such that

- Dataset A had AREAP, ETHGROUP, QUALNUM, SEX, TENURE
- Dataset B had AGE, LTILL, MSTATUS, SEX, TENURE

TENURE and SEX were therefore the overlapping variables. For both dataset A and B a logistic regression was the analytical output, with TENURE being used as the target variable and the remaining variables as predictors.

Run 7 produced the optimal dataset, with the lowest fitness.

Table 7 compares the logistic regression coefficients for output A to those obtained from the synthetic data produced by run 7. compares the logistic regression coefficients for output B to those obtained from the synthetic data produced by run 7

Table 7: Experiment 4.2, regression coefficients for synthetic run 7 compared to the original regression A coefficients

Feature	Original Coefficient	Synthetic Coefficient	Absolute difference
Constant	-0.499	-0.404	0.095
AREAP_58	-0.331	-0.319	0.012
AREAP_59	-0.490	-0.478	0.012
AREAP_60	0.250	0.266	0.016
AREAP_61	-0.793	-0.787	0.007
AREAP_62	0.008	0.020	0.012
AREAP_63	0.109	0.123	0.014
ETHGROUP_10	0.858	0.860	0.003
ETHGROUP_2	0.648	0.654	0.007
ETHGROUP_3	0.769	0.776	0.007
ETHGROUP_4	1.271	1.278	0.007
ETHGROUP_5	-1.646	-1.639	0.007
ETHGROUP_6	-0.878	-0.875	0.003
ETHGROUP_7	-0.124	-0.117	0.008
ETHGROUP_8	0.634	0.635	0.001
ETHGROUP_9	0.491	0.494	0.003
QUALNUM_1	-1.085	-1.071	0.014
QUALNUM_2	-1.621	-1.613	0.007
SEX_2	0.043	0.063	0.020
		MAE:	**0.013**

Feature	Original Coefficient	Synthetic Coefficient	Absolute difference
Constant	0.100	0.099	0.001
AGE	0.005	0.000	0.005
LTILL_2	-0.708	-0.707	0.000
MSTATUS_2	-0.956	-0.957	0.001
MSTATUS_3	-0.466	-0.465	0.001
MSTATUS_4	0.316	0.316	0.000
MSTATUS_5	-0.174	-0.175	0.001
SEX_2	0.024	0.029	0.005
		MAE:	**0.002**

The holdout data was put through the logistic regression models from output A and output B to measure their accuracy in predicting TENURE. These models were generated using the original data. A logistic regression model was then generated using the synthetic data and the holdout data was put through this. The accuracy for all models is in Table 8. The model built on the synthetic dataset outperforms models A and B, both of which were trained on the original. The synthetic data was therefore able to predict better on a holdout dataset than models trained on the two original datasets.

Table 8: Experiment 4.2, logistic regression model accuracy on holdout data

Model	Accuracy
Output A	0.672
Output B	0.678
Synthetic	0.688

5.4.3 Experiment 4.3

Overall, there were 8 variables (AREAP, AGE, ETHGROUP, LTILL, MSTATUS, QUALNUM, SEX, TENURE) – the datasets (n=25178) were split such that

- Dataset A had AREAP, ETHGROUP, QUALNUM, SEX, TENURE
- Dataset B had AGE, AREAP, LTILL, MSTATUS, SEX, TENURE

TENURE, AREAP and SEX were therefore the overlapping variables. For both dataset A and B a logistic regression was the analytical output, with TENURE being used as the target variable and the remaining variables as predictors.

Run 1 produced the optimal dataset, with the lowest fitness. Table 9 compares the logistic regression coefficients for output A to those obtained from the synthetic data produced by run 1. Table 10 compares the logistic regression coefficients for output B to those obtained from the synthetic data produced by run 1.

Table 9: Experiment 4.3, regression coefficients for synthetic run 1 compared to the original regression A coefficients

Feature	Original Coefficient	Synthetic Coefficient	Absolute difference
Constant	-0.499	-0.354	0.145
AREAP_58	-0.331	-0.339	0.008
AREAP_59	-0.490	-0.444	0.046
AREAP_60	0.250	0.287	0.038
AREAP_61	-0.793	-0.784	0.009
AREAP_62	0.008	0.024	0.016
AREAP_63	0.109	0.114	0.005
ETHGROUP_10	0.858	0.860	0.002
ETHGROUP_2	0.648	0.658	0.010
ETHGROUP_3	0.769	0.777	0.008
ETHGROUP_4	1.271	1.275	0.004
ETHGROUP_5	-1.646	-1.637	0.009
ETHGROUP_6	-0.878	-0.869	0.009
ETHGROUP_7	-0.124	-0.114	0.010
ETHGROUP_8	0.634	0.643	0.009
ETHGROUP_9	0.491	0.505	0.014
QUALNUM_1	-1.085	-1.060	0.026
QUALNUM_2	-1.621	-1.608	0.013
SEX_2	0.043	0.100	0.057
		MAE:	**0.023**

Table 10: Experiment 4.3, regression coefficients for synthetic run 1 compared to the original regression B coefficients

Feature	Original Coefficient	Synthetic Coefficient	Absolute difference
Constant	0.154	0.152	0.002
AGE	0.005	0.002	0.003
AREAP_58	-0.345	-0.338	0.007
AREAP_59	-0.362	-0.370	0.008
AREAP_60	0.303	0.297	0.006
AREAP_61	-0.714	-0.711	0.003
AREAP_62	0.068	0.068	0.000
AREAP_63	0.088	0.089	0.001
LTILL_2	-0.691	-0.692	0.001
MSTATUS_2	-0.949	-0.950	0.001
MSTATUS_3	-0.426	-0.426	0.000
MSTATUS_4	0.352	0.352	0.000
MSTATUS_5	-0.164	-0.165	0.002
SEX_2	0.025	0.028	0.002
		MAE:	**0.003**

The holdout data was put through the logistic regression models from output A and output B to measure their accuracy in predicting TENURE. These models were generated using the original data. A logic regression model was then generated using the synthetic data and the holdout data was put through this. The accuracy for all models is in Table 11. The model built on the synthetic dataset outperforms models A and B, both of which were trained on the original. The synthetic data was therefore able to predict better on a holdout dataset than models trained on the two original datasets.

Table 11: Experiment 4.3, logistic regression model accuracy on holdout data

Model	Accuracy
Output A	0.672
Output B	0.679
Synthetic	0.686

6. Conclusions and Future Work

This paper has demonstrated that analytically coherent synthetic datasets can be generated using only cleared analytical outputs released by multiple independent data owners, without any access to underlying microdata. By framing synthesis as a utility-driven optimisation problem and jointly enforcing multiple analytical constraints through an evolutionary algorithm, the proposed framework successfully produces a single, internally consistent synthetic dataset that reproduces key analytical results across a range of distributed scenarios.

Across all experimental settings—from partial variable overlap to increasing numbers of datasets and heterogeneous analytical outputs—the joint synthesis approach consistently achieved low error in reproducing regression coefficients and summary tables. Importantly, it

outperformed alternatives based on local datasets. The results show stable optimisation behaviour, good scalability from two to four datasets, and robustness to differing patterns of overlap in both samples and variables. Minor discrepancies were largely confined to sparse subgroups and did not undermine the interpretability or pedagogical usefulness of the synthetic data.

The findings reinforce the value of output-constrained synthesis as a practical and governance-compatible approach for secure research environments. By relying exclusively on already cleared outputs and non-disclosive metadata, the framework aligns naturally with existing disclosure control processes and avoids introducing new disclosure risks. The emphasis on teaching, evaluation, and methodological development further highlights the suitability of this approach for contexts where analytical fidelity is prioritised over record-level realism.

There are several avenues for future work. First, while this study focused primarily on categorical variables and logistic regression outputs, the framework could be extended to accommodate a broader range of analytical outputs, including continuous outcomes, interaction terms, multilevel models, and time-series summaries. Exploring how competing or partially inconsistent outputs affect convergence and solution quality would also be valuable, particularly in more heterogeneous federated settings.

Second, the current implementation treats all analytical constraints with equal weight. Future research could investigate adaptive or user-specified weighting schemes to reflect differing priorities across outputs or data owners. Relatedly, incorporating explicit measures of uncertainty in released outputs, such as confidence intervals, could allow synthesis targets to be defined as acceptable ranges rather than point estimates.

Third, while EAs proved effective in this study, alternative optimisation strategies—such as hybrid EA–gradient methods or constraint programming approaches—may improve computational efficiency, especially as the dimensionality of the synthetic dataset grows. Formal analysis of convergence properties and scalability limits would further strengthen the methodological foundation.

Finally, future work should consider governance and usability aspects, including how this approach could be embedded into secure research workflows and how synthetic datasets generated in this way are perceived and used by instructors, students, and analysts. Evaluating downstream teaching outcomes and exploratory analyses conducted on these synthetic datasets would provide important evidence of real-world utility.

Overall, this work shows that federated-style synthetic data generation from distributed analytical outputs is not only feasible but effective, offering a transparent, low-risk pathway to producing analytically useful synthetic data in highly restricted access environments.

Acknowledgements

This research was funded by the Economic and Social Research Council under grant number ES/Z502984/1.

References

[1] Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation* (Vol. 201). Springer Science & Business Media.

[2] Back, T., (1996). *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press.

[3] Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468.

[4] Little, R. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics 9(2), 407–426. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf

[5] Pilgram, L., Dankar, F. K., Drechsler, J., Elliot, M., Domingo-Ferrer, J., Francis, P., ... & El Emam, K. (2025). A consensus privacy metrics framework for synthetic data. Patterns, 6(10).

[6] Taub, J., Elliot, M., Pampaka, M., & Smith, D. (2018). Differential correct attribution probability for synthetic data: an exploration. In *International conference on privacy in statistical databases* (pp. 122-137). Cham: Springer International Publishing.

[7] Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of official statistics*, *21*(3), 441.

[8] Nowok, B. (2015). Utility of synthetic microdata generated using tree-based methods. *UNECE Statistical Data Confidentiality Work Session*, 1-11.

[9] Young, J., Graham, P., & Penny, R. (2009). Using Bayesian networks to create synthetic data. *Journal of Official Statistics*, *25*(4), 549-567.

[10] Miletic, M., & Sariyar, M. (2024). Assessing the potentials of LLMs and GANs as state-of-the-art tabular synthetic data generation methods. In *International Conference on Privacy in Statistical Databases* (pp. 374-389). Cham: Springer Nature Switzerland.

[11] Little, C., Elliot, M., Allmendinger, R., & Samani, S. S. (2021). Generative adversarial networks for synthetic data generation: a comparative study. *arXiv preprint arXiv:2112.01925*.

[12] Van Breugel, B., Sun, H., Qian, Z., & van der Schaar, M. (2023). Membership inference attacks against synthetic data through overfitting detection. *arXiv preprint arXiv:2302.12580*.

[13] Elliot, M., Little, C., & Allmendinger, R. (2024). The production of bespoke synthetic teaching datasets without access to the original data. In *International Conference on Privacy in Statistical Databases* (pp. 144-157). Cham: Springer Nature Switzerland.

[14] Little, C., Elliot, M., & Allmendinger, R. (2025, October). Producing synthetic teaching datasets using evolutionary algorithms. In *UNECE Expert Meeting on Statistical Data Confidentiality: Expert meeting on Statistical Data Confidentiality*. https://unece.org/sites/default/files/2025-12/SDC2025_Sd_UnivManchester_SyntheticTeachingDatasets_Elliot_D.pdf

[15] Casaletto, J., Bernier, A., McDougall, R., & Cline, M. S. (2023). Federated analysis for privacy-preserving data sharing: a technical and legal primer. *Annual review of genomics and human genetics*, *24*(1), 347-368.

[16] Li, L., Fan, Y., Tse, M., & Lin, K. Y. (2020). A review of applications in federated learning. *Computers & Industrial Engineering*, *149*, 106854.

[17] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, *14*(1–2), 1-210.

[18] Little, C., Elliot, M., & Allmendinger, R. (2023). Federated learning for generating synthetic data: a scoping review. *International Journal of Population Data Science*, *8*(1), 2158.

[19] Office for National Statistics. Census Division, University of Manchester. Cathie Marsh Centre for Census and Survey Research. (2013). *Census 1991: Individual Sample of Anonymised Records for Great Britain (SARs)*. [data collection]. UK Data Service. SN: 7210, DOI: http://doi.org/10.5255/UKDA-SN-7210-1

Appendices

A1: Data dictionary for UK 1991 SARS Census data

Variable name	Description	Label	Values
AGE	age		
AREAP	Individual SAR area	57	Birmingham
		58	Coventry
		59	Dudley
		60	Sandwell
		61	Solihull
		62	Walsall
		63	Wolverhampton
ETHGROUP	Ethnic group	1	White
		2	Black Caribbean
		3	Black African
		4	Black other
		5	Indian
		6	Pakistani
		7	Bangladeshi
		8	Chinese
		9	Other-Asian
		10	Other-other
LTILL	Limiting long-term illness	1	Yes
		2	No
MSTATUS	Marital status	1	Single
		2	Married
		3	Remarried
		4	Divorced
		5	Widowed
QUALNUM	No. of higher educational qualification	0	None
		1	One
		2	Two
SEX	Sex	1	Male
		2	Female
TENURE	Tenure of household space (derived into binary variable)	0	Own house
		1	Rent house

A2: Regression model reference categories

Variable	Reference category	Label
AREAP	57	Birmingham
ETHGROUP	1	White
LTILL	1	Yes
MSTATUS	1	Single
QUALNUM	0	None
SEX	1	Male