The Impact of Synthetic Data Generation on Data Utility with Application to the 1991 UK Samples of Anonymised Records

TRANSACTIONS ON DATA PRIVACY 13 (2020) 1–23

The Impact of Synthetic Data Generation on Data Utility with Application to the 1991 UK Samples of Anonymised Records

Jennifer Taub*, **Mark Elliot***, **Joseph W. Sakshaug****

*Cathie Marsh Institute, The University of Manchester, Manchester, UK.

**Institute for Employment Research, Ludwig Maximilian University of Munich, and University of Mannheim, Germany.

E-mail: jennifer.taub@.manchester.ac.uk,mark.elliot@manchester.ac.uk, joe.sakshaug@iab.de

Received 1 March 2018; received in revised form 4 January 2019; received in revised form 24 August 2019; received in revised form 28 January 2020; accepted 30 January 2020

Abstract. Synthetic data generation has been proposed as a flexible alternative to more traditional statistical disclosure control (SDC) methods for minimising disclosure risk. However, a barrier to the use of synthetic data is the uncertainty about the reliability and validity of the results that are derived from these data. Surprisingly, there has been a relative dearth of research on how to measure the utility of synthetic data. Utility measures developed to date have been either information theoretic abstractions or somewhat arbitrary collations of statistics, and replication of previously published results has been rare. In this paper, we adopt a methodology previously used by Purdam and Elliot (2007), in which they replicated published analyses using disclosure-controlled versions of the same microdata used in said analyses and then evaluated the impact of disclosure control on the analytic outcomes. We utilise the same studies as Purdam and Elliot, based on the 1991 UK Samples of Anonymised Records, to facilitate comparisons of synthetic data utility between different utility metrics.

Keywords. Synthetic data; CART; multiple imputation; utility metrics

1 Introduction

With the increasing centrality of data in our lives, societies and economies and the drive for greater government transparency and release of open data, there has been a concomitant increase in demand for public release microdata. This demand cannot always be met by using traditional statistical disclosure control (SDC) techniques as SDC techniques can - if they are to provide sufficient confidentiality protection - strongly impact data utility. This is illustrated by Purdam and Elliot (2007) who tested the data utility of two SDC techniques – local suppression, which creates missing values to replace some of the key variables, and Post-Randomization (PRAM) which swaps categories for selected variables based on a predefined transition matrix – and found that these SDC techniques had a significant impact on a sample of published analytic outputs.

An alternative to traditional SDC techniques is synthetic data. The idea of synthetic data was first introduced by Rubin (1993), who proposed multiply imputing an entire dataset, so that no real micro-data would be released. As an alternative, Little (1993) introduced a method that would only replace some of the variables in the observed data, referred to as partially synthetic data. Since fully-synthetic data does not contain any original data, the disclosure of sensitive information from the synthetic data is much less likely to occur. Likewise for partially synthetic data, the sensitive values are synthetic, and thus disclosure of sensitive information is also less likely to occur compared to the original data.

However, to be useful, synthetic data must yield valid statistical analyses. Validity is important since synthetic data can be used to study policy-relevant outcomes and inform policy decisions; for example, the US Census Bureau released a synthetic version of the Longitudinal Business Database (Kinney et al, 2014; 2011) and in Germany a synthetic version of the IAB Establishment Panel has been released (Drechsler et al, 2008a; 2008b). Both of these datasets provide relevant information on businesses in their respective countries, and the quality of these data are of great importance to data users. If the synthetic data produces results that are distorted, incorrect economic conclusions could be drawn. Reiter (2005b) argues that the validity of synthetic data is dependent on the models used to generate them and will not reflect relationships that are present in the original data but not represented in the data generation model. He, also, points out that if the distributional assumptions built into the model are incorrect, then these incorrect assumptions will also be built into the users' analysis models. The importance of this issue is exacerbated by the fact that the vast majority of utility tests for synthetic data are not necessarily representative of the work that data analysts intend to undertake.

Rubin's initial proposal for producing synthetic data was based on multiple imputation (MI) techniques using parametric modelling. Originally, MI was created by Rubin in the 1970s as a solution to deal with missing data by replacing missing values with multiple values, to account for the uncertainty of the imputed values. Recent research has examined non-parametric methods - including machine learning techniques - which are better at capturing non-linear relationships for generating synthetic data (Drechsler and Reiter, 2011). These methods include classification and regression tress (CART) (Reiter, 2005), random forests (Caiola and Reiter, 2010), bagging (Drechsler and Reiter, 2011), support vector machines (Drechsler, 2010), and genetic algorithms (Chen et al, 2016). CART, originally developed by Breiman et al (1984) as a non-parametric modelling tool based on decision trees, has become the most commonly used nonparametric method for generating synthetic data. For example, Kinney et al (2014) used CART to generate synthetic data for some of the variables in the US Longitudinal Business Database.

To generate synthetic data from a CART model, Reiter (2005) explains that each variable is fitted to a tree that splits into branches based on a series of binary splits. These branches continue to divide until they terminate in leaves. The values in the terminal leaf represent the conditional distribution of the predicted variable given the covariates used to grow the tree. The synthetic data is then generated by sampling from the leaves. Reiter (2005) notes the advantages of the CART synthetic data method, including (1) straightforward implementation especially for non-smooth continuous data, and (2) providing a semi-automatic way to fit the most important relationships in the data (p.12). Previous work by Drechsler and Reiter (2011) found that when comparing different methods of machine learning for synthetic data generation, CART yielded the highest data utility. They tested the utility by comparing distributions of different variables and coefficient estimates from a fitted logistic regression model. Likewise, Nowok (2015) compared different tree-based synthetic datasets against a parametric synthetic dataset by using coefficient estimates from a linear

regression, a logistic regression, and a Poisson regression to test data utility. Compared to the other tree-based synthetic data approaches, Nowok found CART performed better, but not as well as the parametric method. Previous studies tend to be limited in their assessment of data utility in that they typically choose only a limited number of analytic models and outcomes, and additionally they tend to analyse said models using a single metric (e.g. confidence intervals).

Synthetic data was originally designed under a multiple imputation framework, however, new rules around the calculation of the variance (Raab et al, 2016) have raised the question of whether a singly imputed synthetic dataset is sufficient to produce high levels of analytical validity. Raghunathan et al (2003) and Reiter et al (2002) introduced the classic combining rules for multiply imputed synthetic data in the early 2000s and for over a decade they were the only set of combining rules available until Raab et al (2016) introduced a new approach for calculating variance estimates for fully synthetic data. Unlike the Reiter/Raghunathan variance equations, the Raab variance equation can produce valid variance estimates for singly imputed fully synthetic data. MI can be very time consuming to produce in comparison to SI. However, the main benefit of MI is that it accounts for the uncertainty of imputations. However, if the Raab variance equation can account for the uncertainty of a single imputation synthesis, this could be potentially a boon to the dissemination of synthetic data. Another argument for the exploration of SI synthetic data is given CART's tendency to overfit, the multiples of CART might not be as diverse as those of traditional parametic MI therefore reducing the benefit of multiple imputation. While Drechsler and Reiter (2009) and Reiter (2002) have explored how the number of imputations has affected parametric synthetic data utility, no such study has been done for CART synthetic data. If CART synthetic data does not have a congruent utility increase alongside a disclosure risk increase, then multiple imputations of CART may not have a beneficial risk-utility trade-off.

In this paper, we focus on assessing how well singly and multiply imputed CART generated synthetic data can replicate real world analyses. Following Purdam and Elliot (2007), we replicate 9 different sets of analyses involving 28 different tests and models to evaluate the utility of CART synthetic data. The remainder of this article is organised as follows. In Section 2 we discuss some alternative definitions of data utility, give a brief overview of data utility tests, and state the research questions addressed in this paper. In section 3 we describe the dataset being utilised and provide an overview of methods used for synthesising and combining the data. Additionally, we describe the methods used to measure synthetic data utility. Section 4 presents the results of our evaluation and discusses the effectiveness of different utility metrics and the differences in SI and MI CART synthetic data, leading to the overall conclusion in Section 5.

2 Data Utility

To describe data utility, Winkler (2005) introduces the terms analytically valid and analytically interesting. He classifies a dataset as analytically valid if the following criteria are approximately preserved:

- Means and covariances on a small set of subdomains
- Marginal values for a few tabulations of data
- At least one distributional characteristic

Winkler (2005) classifies a dataset as analytically interesting if it provides at least six variables on important sub-domains that can be validly analysed. However, both concepts, analytically valid and analytically interesting, do not answer the most important question – will the perturbed dataset produce the same results as the original for analyses that are of most interest to data users? Winkler asserted that it would be impossible to create a file that satisfied a large number of analytic needs while still remaining confidential. A side effect of this proposition is that Winkler's definition of utility does not directly address the needs of the analysts or their research, but rather attempts to address issues of the dataset itself - it is *data-focused*. However, it should be noted that when Winkler's work was published synthetic data was still in its infancy. With the increasing popularity of data synthesis and its potential to produce low risk datasets, a more ambitious perspective on approaching data utility is warranted.

Purdam and Elliot (2007) define the loss of analytical validity as occurring when "a disclosure control method has changed a dataset to the point at which a user reaches a different conclusion from the same analysis" (p. 1102). This definition is more in keeping with the *analysis-focused* approach to data utility, which we also adopt here. While the Winkler definition of data utility is certainly easier to assess than Purdam and Elliot's definition, it could classify data as having high utility but allow for analyses that come to erroneous conclusions.

2.1 Common Utility Measures

Drechsler and Reiter (2009, p.592) classify two existing types of utility measures for synthetic data:

1. Comparisons of broad differences between the original and released data (broad measures)

2. Comparisons of differences in specific models between original and released data (narrow measures)

Traditionally, most synthetic data applications have used narrow measures of data utility. For example, Drechsler et al (2008a) and Drechsler and Reiter (2009) used confidence interval overlap, a narrow measure developed by Karr et al (2006), to test the utility of both fully and partially synthetic data. Alternatively, broad measures tend to quantify some kind of statistical distance between the entire original and released datasets, utilising measures such as the Kullback-Leibler divergence (Karr et al, 2006) or the Hellinger distance (Shlomo et al, 2015). More recently, Snoke et al (2016) adapted Woo et al's (2009) Propensity Score Measure mean-squared error (pMSE), to be compatible with synthetic data, by creating two alternatives to the traditional pMSE: the pMSE ratio and the standardised pMSE (See Snoke et al for more details).

Both broad and narrow utility measures have their weaknesses. Karr et al (2006) reflect that narrow measures are really good for certain analyses, but they do not give a complete picture of a dataset's utility and that broad measures are "pretty good" for many analyses but "really good" for none. In essence we need a utility measure that bridges the gaps between the broad and narrow utility measures, that avoids the traps of being theoretic abstractions, while showing utility for a multitude of data uses.

Purdam and Elliot (2007) introduce a more hands-on approach to evaluating data utility using narrow measures but across a sample of analyses, so as to avoid the trap where narrow measures report artificially high or low utility measures based a single set of analyses.

Their method could be considered a systematically sampled narrow measure, as opposed to a traditional *ad hoc* narrow measure. Using the 1991 Sample of Anonymised Records (SAR) from the British census, they examined a sample of twenty-three published papers, replicated the analyses for ten of them on SDC perturbed data, and compared their results to the original (published) results. This method for evaluating the utility of perturbed data will henceforth be referred to as the Purdam-Elliot methodology. The Purdam-Elliot methodology has the benefit of testing data utility by using analyses that have actually been published and which are of interest to a substantive audience, as opposed to evaluations based on hypothetical analyses (such as those used by Nowok (2015)). Using particular sets of analyses to test data utility has been used in previous studies to evaluate synthetic data (Drechsler et al, 2008b; Lee et al, 2013), however, the Purdam-Elliot methodology provides wider range of applicability due to the volume and variety of analyses that can be used.

In this paper, we use the same base data and same analyses as Purdam and Elliot (2007)[1] (See Appendix A and B for description of the analyses) to evaluate the utility of synthetic data. This approach is beneficial for two reasons: first, the papers that they selected showcase a variety of analytic techniques, including cross-tabulations, logistic regression, probit regression, and multi-level modelling; second, it will allow direct comparisons of the synthetic data results with the SDC controlled data results presented in Purdam and Elliot (2007). This paper will address the following research questions:

1. How does CART synthetic data perform when using the Purdam-Elliot methodology?

2. How does the Purdam-Elliot methodology compare to numerical utility metrics in its application to synthetic data?

3. Does singly versus multiply imputed CART synthetic data affect utility?

3 Data Sources and Methods

Following Purdam and Elliot (2007), we use the 1991 individual SAR as our base dataset. This dataset consists of 1,116,181 records, which represents a 2% sample of the population of Great Britain. The SAR is available to researchers under an end user license and contains information on topics such as age, gender, ethnicity, household size, household type, employment, and health[2]. We utilize three versions of these data: 1) the original 1991 SAR without any perturbation as a control; 2) ten singly imputed (SI) CART-generated synthetic versions of the 1991 SAR and; 3) a multiply imputed (MI) CART-generated synthetic version of the 1991 SAR with m=10.

We use ten different singly imputed synthetic datasets, to reduce the risk of drawing a very good SI dataset by chance. For each singly imputed dataset the utility metrics were calculated separately.We then averaged the results across the 10. Realistically, if a data producer were to be using single imputation then their goal would be to release only one such dataset, and therefore an analyst would only have one dataset with which to run their analyses. However, given that this is an empirical test, it is more robust to average across multiple draws.

[1] Of the original ten papers that were reanalysed in Purdam and Elliot (2007), only nine were used for this paper due to problems with replicating the 10th.

[2] The 1991 SAR can be found at https://www.ukdataservice.ac.uk

3.1 Creation of Synthetic Data Files

CART is used to synthesise each variable of the SAR data (see appendix C for synthesising order and appendix D for descriptions of the variables). Due to the size of the dataset and the subsequent computational load, the data was randomly divided into 100 subsets and each subset was synthesized independently for both *m*=1 and *m*=10. Then the 100 samples were aggregated together to form the entire synthetic dataset.

The CART synthetic data were generated using the r-package *synthpop* version 1.3-0 (Nowok et al, 2016). The method for *synthpop* was set to "CART", which is derived from the *rpart* package. For comparison, the original (unperturbed) SAR is used as a control to ensure that all analyses are correctly replicated. The evaluation considers both a singly imputed CART synthetic dataset (*m*=1) and a multiply imputed CART synthetic dataset (*m*=10). The missing data from the original 1991 SAR are left unchanged in the synthesis process and no attempt is made to impute them. We used the default setting of *synthpop*, which included *proper=FALSE*, which means the synthetic data was generated without drawing from the posterior distribution. This is important to note since it can lead to some bias when using the Ragunathan et al (2003) and Reiter (2002) combining rules described in 3.1.1.

3.1.1 Imputation Combining Rules

To analyse the MI CART synthetic datasets, the combining rules from Ragunathan et al (2003) were employed, along with the variance equations from Reiter (2002) and Raab et al (2016).

A point estimate of the target parameter derived from the synthetic data can be obtained through the following expression:

$$\bar{q}_m = \frac{1}{m}\sum_{i=1}^{m} q_i \text{ with } i = 1, ..., m \tag{1}$$

where $\bar{q}_m$ is the estimated mean of the individual estimates q_i derived from each of the m synthetic populations.

Ragunathan et al calculate the total variance of the point estimate as follows:

$$T_s = (1 + m^{-1})b_m - \bar{v}_m \tag{2}$$

where $\bar{v}_m$ refers to the mean of the within-imputation variances and b_m refers to the between-imputation variance. The mean of the within-imputation variance is calculated as:

$$\bar{v}_m = m^{-1}\sum_{i=1}^{m} v_i \tag{3}$$

where v_i is the estimate of variance of the point estimate based on synthetic dataset i. The between-imputation variance is calculated as:

$$b_m = \frac{1}{m-1}\sum_{i=1}^{m}(q_i - \bar{q}_m)^2 \tag{4}$$

Reiter (2002) notes that using equation 2 to calculate the total variance can lead to a negative variance and proposed the following fix to account for this problem:

$$T_s^* = \max(0, T_s) + \delta * (\frac{n_{syn}}{n}\bar{v}_m) \tag{5}$$

where $\delta = 1$ if $T_s < 0$, and $\delta = 0$ otherwise. This means that if the variance from equation 2 is positive it would not be altered and if negative it would be adjusted. As yet another way of estimating variance for synthetic data, Raab et al (2016) suggested the following:

$$T_s = \bar{v}_m(1 + \frac{1}{m}) \tag{6}$$

With the Raab combining rules $\bar{v}_m$is intended to account for both the between imputation variance and the within imputation variance, therefore b_m is omitted from equation 6. Since b_m is not included in equation 6, there is nothing inherently specific to multiple imputation in equation 6, and therefore one can obtain valid variance estimates for singly-imputed synthetic datasets. This is in contrast to equations 2 and 5, which are specifically designed for making inference from multiply imputed synthetic datasets. It should be noted that the Raab et al combining rules were created for data derived from a simple random sample (SRS). The SAR is not strictly a SRS, but rather a stratified random sample. However, the stratification is minimal and we concur with Marsh (1993) who argues that the 2% individual SAR can be treated as a SRS.

In this article, we calculate point estimates for the CART m=10 synthetic data using equation 1 and calculate the variance using both equations 5 and 6. For the singly imputed CART synthetic data we calculate the variance using equation 6. We have selected to use both sets of combining rules since they have different advantages. The Reiter/Ragunathan combining rules are more established in the synthetic data literature. However, Raab et al (2016) makes a clear case that the previous rules are not needed and there is the additional advantage that the Raab rules can be used for SI synthetic data, allowing a direct comparison between MI and SI synthetic data. Since one of the utility metrics that we employ is based on overlapping confidence intervals, the way in which the variance is estimated can potentially make a significant difference in the evaluation. Thus, it's important to compare the impact of the different combining rules.

3.2 Utility Metrics

We assess the utility of synthetic data using three measures: confidence interval overlaps, ratio of estimates, and severity ratings. To calculate the confidence interval overlap (CIO) we use 95% confidence intervals. The CIO is calculated as:

$$J_k = \frac{1}{2}(\frac{U_{,k} - L_{,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{,k} - L_{,k}}{U_{syn,k} - L_{syn,k}}) \tag{7}$$

where $U_{,k}$ and $L_{,k}$, denote the respective upper and lower bounds of the intersection of the confidence intervals from both the original and synthetic data for estimate k, $U_{org,k}$ and $L_{org,k}$ represent the upper and lower bounds of the original data, and $U_{syn,k}$ and $L_{syn,k}$ of the synthetic data. The CIO is then averaged across all coefficients in the model to give an overall mean for each model. Likewise, for tabular data all cells are averaged to give a mean CIO. Separate CIOs are calculated for each of the different combining rules used for variance estimation.

The ratio of estimates (ROE) is a new measure that is calculated by taking the ratio of the synthetic and original data estimates, where the smaller of these two estimates is divided by the larger one. Thus, given two corresponding estimates (e.g. totals, proportions), where y^1_{orig} is the estimate from the original data and y^1_{synth} is the corresponding estimate from

Severe	The results are sufficiently different
Moderate	Change in emphasis rather than a completely different finding
No effect	Indicates that the figures may be slightly different but the overall pattern is not, indicating the same conclusion can be consistently drawn

Table 1: Key for Severity Ratings

	A	B
C	0.95	0.05
D	0.10	0.90

Original

	A	B
C	0.97	0.03
D	0.15	0.85

No Effect

	A	B
C	0.70	0.30
D	0.20	0.80

Moderate

	A	B
C	0.55	0.45
D	0.40	0.60

Severe

Figure 1: An Example of the Paper could Effect Severity

the synthetic data, the ROE is calculated as:

$$\frac{\min(y^1_{orig}, y^1_{synth})}{\max(y^1_{orig}, y^1_{synth})} \tag{8}$$

If $y^1_{orig} = y^1_{synth}$ then the ROE = 1. Given that regression models show more complicated results, such as coefficients with positive and negative values, we will be restricting the ROE to cross-tabs and frequency tables. This allows the ratio of estimates to provide a value between 0 and 1. For each categorical variable the ratio of estimates are averaged across categories to give an overall ratio of estimates for each synthetic variable. We refer to both ROE and CIO as utility metric scores.

In contrast to the utility metrics described above, the severity ratings take a more holistic approach to assessing data utility. Using the key in Table 1, which is taken from the definitions provided by Purdam and Elliot (2007, p. 1109), each synthetic data analysis is evaluated on how severely[3] the analysis is affected by the synthesis. The severity rating is not based upon the findings of CIO or ROE scores, but rather the conclusions of each paper are taken into account to assess whether or not the authors of the original papers would come to the same conclusions using synthetic/perturbed data as they had in their papers using the original data. Figure 1 illustrates an example of what would constitute the different severity effects, given that the analyst's had concluded that condition AC $>$ BC and AD $<$ BD. As shown, the "no effect" classification does not require the perturbed dataset to be identical with the original but the values are sufficiently similar that they would not affect the conclusions drawn. The moderate effect example has more markedly different numbers, but shows a similar overall relationship between the two variables as the original. Finally, in the severe effect example, the relationship found in the original data is not respected.

4 Results & Discussion

Table 2 shows the severity ratings, average ROE and CIO scores for the singly imputed and multiply imputed CART synthetic data. For more detail, Table 3 breaks the utility metrics

[3] "Severity" is a subjective measure based on the authors' judgement.

down by table for the papers based on cross-tabulations/frequency tables and Table 4 does the same for the regression models. Table 3 shows that the ratio of estimates and confidence interval overlaps can vary drastically from one another, while Table 3 and 4 show that for MI synthetic data the Reiter and Raab methods for calculating the variance lead to similar CIO scores. Appendix E shows the results for the individual single imputation scores.

Table 5 summarises the severity ratings from Table 2, and also includes the findings from Purdam and Elliot on suppression and PRAM. The singly imputed CART synthetic data has lower utility than the multiply imputed CART synthetic data. The MI data performs similarly to PRAM and better than when the data was perturbed by both suppression and PRAM. Neither synthetic dataset performs better than suppression alone.

Paper	Severity Rating SI	ROE SI	CIO Raab SI	Severity Rating MI	ROE MI	CIO Reiter	CIO Raab MI
Ballard (1996)	Moderate	0.704	0.408	Moderate	0.685	0.36125	0.330
Champion (1996)	Moderate	0.782	0.573	Moderate	0.835	0.519	0.456
Drinkwater and OLeary (1997)	Moderate	NA	0.569	Moderate	NA	0.611	0.616
Eade et al (1996)	Severe	0.785	0.349	Severe	0.805	0.386	0.372
Gardiner and Hill (1996)	Moderate	NA	0.594	Moderate	NA	0.638	0.64
Gardiner and Hill (1997)	Severe	0.746	0.484	Moderate	0.673	0.418	0.352
Gould and Jones (2000)	Severe	NA	0.191	Moderate	NA	0.554	0.546
Green (1997)	Moderate	0.771	0.231	Moderate	0.780	0.239	0.270
Leventhal (1994)[4]	No effect	0.810	NA	No effect	0.839	NA	NA

Table 2: Severity Rating, Ratio of Estimates, and CIO for the Overall Paper

[4]For Tables 2 and 3 Leventhal was left blank for the CIO due to the complexity of how the figure was calculated (See Leventhal (1994) for more details).

Paper	Table or Figure	Severity Rating SI	ROE SI	CIO Raab SI	Severity Rating MI	ROE MI	CIO Reiter	CIO Raab MI
Ballard (1996)	Figure 5.1	No effect	0.618	0.502	No effect	0.585	0.449	0.388
	Figure 5.2	Moderate	0.594	0.531	Moderate	0.545	0.457	0.430
	Figure 5.3	Moderate	0.784	0.333	Moderate	0.789	0.304	0.280
	Figure 5.4	Moderate	0.818	0.265	Moderate	0.822	0.235	0.221
Champion (1996)	Table 4.7	Moderate	0.782	0.573	Moderate	0.835	0.519	0.456
Eade et al (1996)	Table 6.3	Moderate	0.789	0.343	Moderate	0.795	0.319	0.337
	Table 6.4	Severe	0.780	0.354	Severe	0.786	0.319	0.322
Gardiner and Hill (1997)	Table 1	Severe	0.795	0.346	Severe	0.646	0.278	0.263
	Table 2	Moderate	0.804	0.508	Severe	0.644	0.354	0.340
	Table 3	Moderate	0.723	0.482	Moderate	0.579	0.313	0.292
	Table 4	Severe	0.660	0.599	No effect	0.823	0.726	0.511
Green (1997)	Table 4.1	Severe	0.742	0.206	Severe	0.748	0.208	0.204
	Table 4.2	Moderate	0.730	0.212	Moderate	0.749	0.202	0.219
	Table 4.3	Severe	0.775	0.190	Moderate	0.781	0.229	0.229
	Table 4.4	Moderate	0.772	0.173	Moderate	0.786	0.189	0.246
	Table 4.5	No effect	0.831	0.373	No effect	0.817	0.357	0.396
	Table 4.6	No effect	0.773	0.233	No effect	0.799	0.246	0.328
Leventhal (1994)	Figure 1	No effect	0.886	NA	No effect	0.891	NA	NA
	Figure 2	No effect	0.760	NA	No effect	0.651	NA	NA
	Figure 3	No effect	0.847	NA	No effect	0.858	NA	NA
	Figure 4	Moderate	0.728	NA	Moderate	0.782	NA	NA
	Figure 5	Moderate	0.716	NA	Moderate	0.777	NA	NA
	Figure 6	No effect	0.890	NA	No effect	0.966	NA	NA
	Figure 7	No effect	0.844	NA	No effect	0.945	NA	NA

Table 3: Severity Ratings, ROE, and CIO for Individual Tables and Figures

Paper	Table or Figure	Severity Rating SI	CIO Raab SI	Severity Rating MI	CIO Reiter	CIO Raab MI
Drinkwater and OLeary (1997)	Table 5 for males	Moderate	0.524	Moderate	0.600	0.606
	Table 5 for females	Moderate	0.614	Moderate	0.622	0.626
Gardiner and Hill (1996)	Table 4	Moderate	0.594	Moderate	0.638	0.640
Gould and Jones (2000)	Table 4 A	Severe	0.191	Moderate	0.554	0.546

Table 4: Severity Ratings and CIO for Individual Regression Models

	No Effect	Moderate	Severe
CART m=1	1	5	3
CART m=10	1	7	1
Suppressions	5	5	0
Post-randomisation	2	7	1
Both Suppressions and PRAM	1	5	4

Table 5: Summary of the Effect of Perturbations (Information on PRAM and Suppressions from Purdam and Elliot, 2007, p. 1110)

4.1 Relationships Between Severity Ratings and Utility Metrics

To determine if there is a relationship between the severity ratings and the utility metrics, six one-way analyses of variance (ANOVA) were conducted. The results are displayed in Table 6. None of the results have a significant F-value at the $p<0.05$ level. However, the CIO score calculated using the Reiter method for the multiply imputed data is significant at the $p<0.1$ level. This shows that there is not a clear relationship between the utility metrics and how well the synthetic dataset performs for a given analysis.

Single Imputation	ROE	F(2,21)= 1.664	P= 0.213
	CIO-Raab	F(2,14)= 0.122	P= 0.886
Multiple Imputation	ROE	F(2, 21)= 1.754	P= 0.197
	CIO-Reiter	F(2,14)= 2.89	P= 0.089
	CIO-Raab	F(2,14)= 1.822	P= 0.198

Table 6: ANOVA between Severity Ratings and Utility Metrics

4.2 Comparing MI and SI CART Synthetic Data

Our third research question concerns whether the multiply imputed CART data performs better than the singly imputed CART data. To answer this question, all utility metrics were utilised. Table 5 shows that the MI CART synthetic data performs slightly better when using the severity ratings, since fewer analyses were severely affected. In comparison, a Pearson's test of the ratio of estimates for the MI and SI synthetic has a correlation of r=0.547, showing that the results are correlated. The confidence interval overlaps for SI and MI CART data show a stronger Pearson's correlation than the ROE (as shown in Table 7). Table 7 additionally shows the Pearson correlation comparing the different combining rules for MI synthetic data. It shows that the Reiter and Raab combining rules result in very similar results.

Overall, the utility metrics show that there is a strong relationship between the utility of SI and MI CART synthetic data. In fact the only paper with major differences was Gould and Jones (2000), which was classified as severe for the SI CART data but moderate for the MI CART data. Overall, we did not see a huge increase in utility for the MI compared to the SI. While there is the concern that SI could be affected by randomness, Appendix E shows the variance between the different SI draws and overall the SI synthetic dataset appears to perform similarly to one another. In fact, for some of the analyses the SI dataset average outperformed the MI dataset, such as the Gardniner and Hill (1997) paper, which shows higher utility scores for the SI synthetic data than for the MI synthetic data.

	CIO-Reiter	CIO- Raab MI
CIO Raab SI	0.777	0.697
CIO- Reiter		0.928

Table 7: Pearson's R Comparing the CIO scores

4.3 Weaknesses with the Utility Metrics

We found that all three utility metrics have weaknesses. One weakness that all three measures share is that they are descriptive measures. While the ROE and Severity Ratings are

more clearly descriptive, the CIO is also descriptive since it merely describes the variance of the measure. There are no measures of data utility currently in the literature that are intended for non-descriptive statistics.

4.3.1 Confidence Interval Overlap

While the confidence interval overlap is the standard method for determining utility it also suffers from two weaknesses.

First, assessing synthetic data utility is not the primary purpose of confidence intervals. The primary purpose of confidence intervals is to make inferences from samples to a population. Synthetic data are not a sample of a population; the data generating process is the synthesis model. When two samples are drawn from the same population, as their sample size increases their confidence intervals would get smaller, but any estimands will also tend to converge monotonically and therefore the confidence interval overlaps would tend to be stable across sample sizes. However, that is not necessarily the case for a synthetic dataset. In this case we would expect that as the dataset size increases any estimand will converge to the model not to the population value. To the extent that the model fails to capture all of the variance in the estimand of interest in the original data then as the dataset size increases any estimand derived from the synthetic data will converge to the population at a slower rate than the rate at which the confidence interval shrinks. Therefore, CIO's tend to worsen as dataset size increases.

Second, the confidence interval overlap score does not distinguish between instances of non-overlap, despite the fact that the degree of non-overlap can vary greatly. Figure 2 displays two examples of non-overlap; image A represents two intervals that just miss each other, while image B shows two intervals that are vastly different. The confidence interval overlap has no way of differentiating near misses versus far misses. All scores that do not overlap are treated the same. One solution would be to use a larger confidence interval such as a 99% confidence interval instead of a 95% confidence interval. It is even possible to double or triple the existing confidence interval so that all scores will overlap. However, this risks causing the opposite problem where all CIO scores are artificially high.

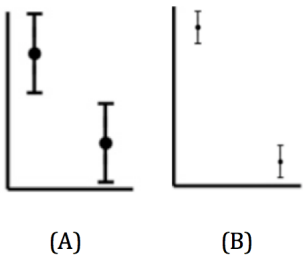
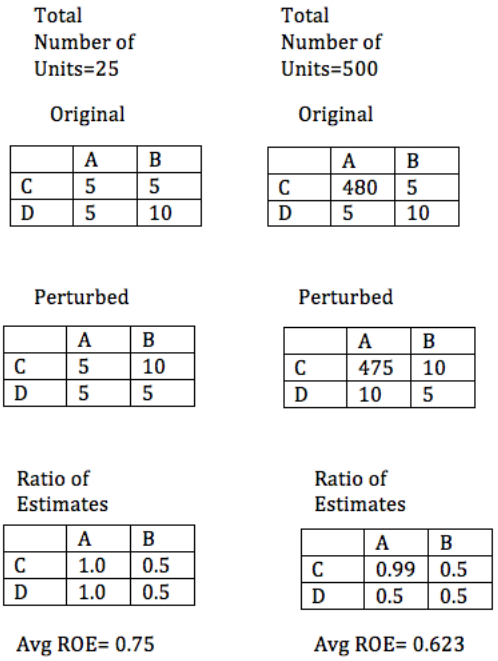


Figure 2: Examples of Confidence Interval Misses

4.3.2 Ratio of Estimates

The ratio of estimates can be problematic for three reasons.

First, in its treatment of zeros. If there is a zero in the numerator, the ratio of estimates gives a value of zero, regardless of the denominator; but intuitively and arithmetically 0 and 1 are much closer than 0 and 1000. Likewise, the ratio of estimates does not take scale into account. For example, if the synthetic data had a score of 1.5% and the original 1%,



	A	B
C	5	5
D	5	10

	A	B
C	480	5
D	5	10

	A	B
C	5	10
D	5	5

	A	B
C	475	10
D	10	5

	A	B
C	1.0	0.5
D	1.0	0.5

	A	B
C	0.99	0.5
D	0.5	0.5

Figure 3: An Example of Contextualization for ROE

the ratio of estimates would be 0.666. However if the synthetic was 91.5% and the original was 91% the ratio of estimates would be 0.995. Given that in many instances a data analyst would consider 1% versus 1.5% quite similar, it seems arbitrary that example 1 has a low ROE score while example 2 is quite high.

Second, the ROE needs contextualisation on whether it is performing well or not. Depending on the number of units in a given table, whether or not the score is informative varies. Figure 3 gives the example of two different perturbed datasets where one has a larger number of units than the other. For both datasets, the ROE for column B is 0.5. However, in the example where the number of units is 500, the difference between 5 and 10 is not that important since it is a very large dataset and 5 and 10 are comparatively both very low counts. In the example where the number of units is 25, the difference between 5 and 10 could be considered more important. However, the example with 25 units actually has a larger ROE score. So, despite the similar average ROE scores, the example with more units is performing better.

Finally, the ROE does not take the uncertainty of the estimator into account. The ability to take uncertainty into account is the main advantage of the CIO as a utility metric. The ROE is on the other hand a more intuitive heuristic, and therefore more appropriate as an initial test for data utility, before using more complicated and time-consuming tests.

4.3.3 Severity Rating

The severity rating is a subjective measurement and therefore cannot be replicated in the same way that the other measures can. Due to this subjectivity, it was - for some analyses - harder to evaluate than for others, where the effect was more obvious.

For example, for the Champion (1996) paper the CART synthetic data follows the relationship that the white ethnic group has more people moving long distances than other ethnic groups, and that black ethnic groups have the lowest numbers moving 200+ km or more, as discussed in the original paper. However, it does not pick up on the finding that more Chinese move 200+ km than other non-white ethnic groups, which was also discussed in the original paper. Since the synthetic data overall has the same findings as the original paper, even with the exception of the Chinese migrant data, we classed it as "moderate." Whilst there is an argument to be made for a severity rating of "no effect", we ultimately decided that picking up on the relationships of the smaller ethnic groups did matter, hence it was classed as moderate.

The Eade et al (1996) paper threw up similar difficulties in classification, being on the boundary between "severe" and "moderate". The Eade et al paper is about the Bangladeshi experience in the UK, although the frequency tables created from the 1991 SAR include all 10 ethnic groups, as well as a category for those born in Ireland. For many of the ethnic groups, the trends shown by the CART synthetic data are very similar to that of the original data, though not for the Bangladeshi ethnic group, which was the focus of their paper. Ultimately, we rated one of the tables as having a "moderate" effect since it captured most of the relationships, though not the focus of the paper, and the other table as "severe" since there was more smoothing of the racial differences. This translates to the Eade paper having a high average ratio of estimates, but a severe effect. One solution to this problem is to add more categories; such as, "no effect", "slight effect", "moderate effect", "moderately severe effect" and "severe effect". However, judging which of the five categories an impact falls into may be just as problematic. Furthermore, the severity ratings are non-trivial to determine and resource intensive.

The severity ratings do however have some advantages compared to the numeric utility metrics. The severity ratings are the only method that takes into account intention. Many government agencies use data that has undergone some form of perturbation and based on these data, important policy decisions can be made. Only a data analyst can go through a perturbed dataset and determine whether the same decisions are likely to be made.

4.4 Analyses Associated with High and Low Utility

The results also raise the question of why the synthetic data produced high utility outputs for the analyses from some papers but not others, and if there are certain kinds of analyses that are more suited to be conducted with synthetic data.

One example of an analysis not replicated was in the Gardiner and Hill (1997) paper. While the sample size for Gardiner and Hill's Table 4 was 3,532, it ran into problems because most of that sample consisted of white participants and they were looking at racial variation. Gardiner and Hill (1997) report that 4.08% of black people in Leicester cycle to work, however that statistic is made up of only two cycling black Leicester residents who make up only 0.000179% of the entire data set. This micro-sample was not replicated by the single imputation synthetic data, but it was by the MI synthetic data (showing some of the advantages gleaned by MI). However, had it not been replicated it would not be regarded as problematic because Gardiner and Hill's reporting of it in the first place was somewhat misleading. In general, we should not be concerned about replicating findings that appear to be spurious.

A paper like Leventhal (1994) was easier for the synthetic datasets to replicate. He was interested in how many people fit a target demographic (head of household, aged 55+, employed, owns home). This was a large category consisting of 1.46% of the overall pop-

ulation and he only ever compared his target variable to one other variable at a time. So when looking at the prevalence of his target demographic in different ethnic groups and so forth the results of the synthetic data stayed true to that of the original.

4.5 Comparisons to other work–Severity Ratings

The SI CART synthetic data (Table 5) shows similar utility to when the data is perturbed by both local suppression and PRAM. The MI CART performed better than the SI data and better than when both local suppression and PRAM were used. It performed similarly to the PRAM dataset, but not as well as the data that was suppressed. Based on their findings, Purdam and Elliot determined that suppression and PRAM had a high level of distortion of the data and that many of the original analyses could not be replicated. As a caveat for these findings we note that all are dependent on the parameter setting for the method and cannot properly be calibrated without also considering disclosure risk. Furthermore, the utility of CART synthetic data is highly dependent on the selection of the complexity parameter. Given that the default value for the complexity parameter in *synthpop* 1.3 is 0.01, this will lead to synthetic data with less analytical validity compared to using a smaller parameter value. Therefore the fact that the synthetic data perform similarly to other SDC methods despite using the default value is a good sign for the utility of the synthetic data.

4.6 Disclosure Risk and Synthetic Data

The focus of this paper is on the utility of synthetic data, however it is worth mentioning disclosure risk. Synthetic data is, in theory, without disclosure risk since its values do not reflect real individuals. However, several papers (Taub et al, 2018; Reiter et al, 2014; McClure and Reiter, 2012; Reiter and Mitra, 2009) have explored the quantification of disclosure risk for synthetic data. These papers have found that synthetic data is not free of disclosure risk. Elliot (2014), followed by Taub et al (2018), used *Differential Correct Attribution Probability* (DCAP) to evaluate synthetic data coming from the *synthpop* package, (the package we use here). Elliot (2014) found that the data produced by *synthpop* contained little disclosure risk. Taub et al found that while the synthetic data files had less disclosure risk than had the original been released, they also found that as the number of imputations increased, so did the the disclosure risk.

5 Conclusion

This paper introduced the Purdam-Elliot methodology as a device for assessing the utility of synthetic data. The Purdam-Elliot methodology has the advantage of running a broad sample of real analyses and striking a balance between the non-specificity of general metrics and the ad hoc nature of traditional narrow measures (since for a particular set of analyses a synthetic dataset may perform with high utility, it may still falter for another analyses, as demonstrated by the results herein). The Purdam-Elliot methodology takes into account that researchers use the data in unexpected ways, by subsetting the data and/or creating their own compilation variables and that it is important to have synthetic data that fulfils these diverse needs within reason. Therefore, this methodology was useful in investigating which kinds of analyses were more amenable to being replicated with synthetic data.

In terms of MI and SI CART synthetic data, the MI data did perform better for some analyses, but not for all. Depending on how complicated an analysis is, a researcher could

be justified in using SI synthetic data. One of the main arguments for using MI is that MI accounts for uncertainty, which is demonstrated by the increased variance. Therefore, the Raab combining rules allow for SI synthetic data to account for the additional variance without the hassle of MI is a great boon. Additionally, the CIO scores did tend to be similar for MI and SI.

Further research should be used to determine if the CIO is correlated more strongly with the severity ratings for regression models. The sample of papers used in this paper favoured frequency tables and therefore the sample of regression models was small. It is possible that there might be a stronger relationship between CIO scores and severity ratings for regression models. Likewise, the ROE scores appeared more indicative of the severity ratings for the frequency tables.

References

[1] Ballard, R. (1996). The Pakistanis: stability and introspection. In: C. Peach, ed., Ethnicity in the 1991 Census: Volume 2 The ethnic minority populations of Great Britain, 1st ed. London: HMSO, pp.121-149.

[2] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees. 1st ed. Belmont, California: Wadsworth, Inc.

[3] Caiola, G. and Reiter, J. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. Transactions on Data Privacy, 3, pp.27-42.

[4] Chen, Y., Elliot, M. and Sakshaug, J. (2016). A Genetic Algorithm Approach to Synthetic Data Production. PrAISe '16 PROEeedings of the 1st International Workshop on AI for Privacy and Security, (13).

[5] Drechsler, J. (2010). Using Support Vector Machines for Generating Synthetic Datasets. Lecture Notes in Computer Science, 6344, pp.148-161.

[6] Drechsler, J. and Reiter, J. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic data. Computational Statistics and Data Analysis, 55, pp.3232-3243.

[7] Drechsler, J. and Reiter, J.P. (2009). Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB Establishment Survey. Journal of Official Statistics, 25(4), 589-603.

[8] Drechsler, J., Bender, S., and Rssler, S. (2008a). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. Transactions in Data Privacy, 1, 105-130.

[9] Drechsler, J., Dundler, A., Bender, S., Rssler, S. and Zwick, T. (2008b). A new approach for disclosure control in the IAB establishment panel–multiple imputation for a better data access. AStA Advances in Statistical Analysis, 92, pp.439-458.

[10] Drinkwater, S. and O'Leary, N. (1997). Unemployment in Wales: Does Language Matter?. Regional Studies, 31(6), pp.583-591.

[11] Eade, J., Vamplew, T. and Peach, C. (1996). The Bangladeshis: the encapsulated community. In: C. Peach, ed., Ethnicity in the 1991 Census: Volume 2 The ethnic minority populations of Great Britain, 1st ed. London: HMSO, pp.150-160.

[12] Gardiner, C. and Hill, R. (1997). Cycling on the Journey to Work: Analysis of the Socioeconomic Variables from the UK 1991 Population Census Samples of Anonymised Records. Planning Practice & Research, 12(3), pp.251-261.

[13] Gardiner, C. and Hill, R. (1996). Analysis of Access to Cars from the 1991 UK Census Samples of Anonymised Records: A Case Study of Elderly Population of Sheffield. Urban Studies, 33(2), pp.269-281.

[14] Green, A. (1997). Patterns of ethnic minority employment in context of industrial and occupational growth and decline. In: V. Karn, ed., Employment, Education and Housing Among the Ethnic Minority Populations of Britain, 1st ed. London: The Stationary Office, pp.67-90.

[15] Gould, M. and Jones, K. (1996). Analyzing perceived limiting long-term illness using U.K. census microdata. Social Science & Medicine, 42(6), pp.857-869.

[16] Karr, A., Kohnen, C., Oganian, A., Reiter, J. and Sanil, A. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician, 60(3), pp.224-232.

[17] Kinney, S., Reiter, J. and Miranda, J. (2014). SynLBD 2.0: Improving the synthetic Longitudinal Business Database. Statistical Journal of the IAOS, 30, pp.129-135.

[18] Kinney, S., Reiter, J., Reznek, A., Miranda, J., Jarmin, R. and Abowd, J. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review, 79(3), pp.362-384.

[19] Lee, J., Kim, I. and O'Keefe, C. (2013). On Regression-Tree-Based Synthetic Data Methods for Business Data. Journal of Privacy and Confidentiality, 5(1), pp.107-135.

[20] Leventhal, B. (1994). "Case Study Examples to Demonstrate the use of Samples of Anonymised Records in Marketing Analysis", Occasional Paper 5, CMU, University of Manchester, Manchester.

[21] Little, R. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9(2), pp.407-426.

[22] Mateo-Sanz, J., Domingo-Ferrer, J. and Sebe, F. (2005). Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata. Data Mining and Knowledge Discovery, 11, pp.181-193.

[23] McClure, D. and Reiter, J. (2012). Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data. Transactions on Data Privacy, 5, pp.535-552.

[24] Nowok, B., Raab, G. and Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), pp.1-26.

[25] Nowok, B. (2015). Utility of synthetic microdata generated using tree-based methods. In: UNECE Statistical Data Confidentiality Work Session.

[26] 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, Accessed 29 June 2017. SN: 7210, http://doi.org/10.5255/UKDA-SN-7210-1

[27] Purdam, K. and Elliot, M. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environment and Planning A, 39, pp.1101-1118.

[28] Raab, G., Nowok, B. and Dibben, C. (2016). Practical data synthesis for large samples. Journal of Privacy and confidentiality. [online] Available at: http://arxiv.org/abs/1409.0217 [Accessed 15 Mar. 2017].

[29] Raghunathan, T., Reiter, J. and Rubin, D. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19(1), pp.1-16.

[30] Reiter, J. (2005a). Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, 21, pp.441-462.

[31] Reiter, J. (2005b). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. Journal of the Royal Statistical Society: Series A (Statistics in Society), 168(1), pp.185-205.

[32] Reiter, J. (2002). Satisfying Disclosure Restriction with Synthetic Data Sets. Journal of Official Statistics, 18(4), pp.181-188.

[33] Reiter, J. and Mitra, R. (2009). Estimating Risks of Identification Disclosure in Partially Synthetic Data. The Journal of Privacy and Confidnetialality, 1(1), pp.99-110.

[34] Reiter, J., Wang, Q. and Zhang, B. (2014). Bayesian Estimation of Disclosure Risks for Multiply Imputed, synthetic Data. Journal of Privacy and Confidentiality, 6(1), pp.17-33.

[35] Rubin, D. B. (1993). Statistical Disclosure Limitation. Journal of Official Statistics, 9(2), 461-468.

[36] Rubin, D. (1987). Multiple imputation for nonresponse in surveys. 1st ed. Hoboken, N.J: John Wiley & Sons.

[37] Rubin, D. (1981). The Bayesian Bootstrap. The Annals of Statistics, 9(1), pp.130-134.

[38] Shlomo, N., Antal, L. and Elliot, M. (2015). Measuring Disclosure Risk and Data Utility for Flexible Table Generators. Journal of Official Statistics, 31(2), pp.305-324.

[39] Snoke, J., Raab, G., Nowok, B., Dibben, C. and Slavkovic, A. (2016). General and specific utility measures for synthetic data.

[40] Taub, J., Elliot, M., Pampaka, M. and Smith, D. (2018). Differential Correct Attribution Probability for Synthetic Data: An Exploration. J. Domingo-Ferrer and F. Montes (Eds.): PSD 2018, LNCS 11126

[41] Winkler, W. (2005). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Research Report Series, 2005(09).

[42] Woo, M., Reiter, J., Oganian, A. and Karr, A. (2009). Global Measures of Data Utility for Microdata Masked of Disclosure Limitation. The Journal of Privacy and Confidentiality, 1(1), pp.111-124.

Appendices

A Description of Analyses used in Papers

Paper	Variables Used	Table	Description
Ballard (1996)	AGE, COBIRTH, DCOUNTY, ETHGROUP, MSTATUS, REGIONP, SEX	Figure 5.1	Bar chart of country of origin and age
		Figure 5.2	Bar char of age and marital status
		Figure 5.3	Pie chart of region
		Figure 5.4	Bar chart of region and ethnicity
Champion (1996)	DISTMOVE, ETHGROUP	Table 4.7	Frequency table of distance moved by ethnicity
Drinkwater and O'Leary (1997)	AGE, CESTSTAT, COBIRTH, DCOUNTY, DEPCHILD, ECOPRIM, MSTATUS, REGIONP, SEX, WELSHLAN	Table 5 males	Probit regression of whether employed by age, marital status, long-term limiting illness, qualifications, housing status, area, and welsh fluency
		Table 5 females	
Eade et al (1996)	AGE, COBIRTH, ETHGROUP, SEX, SOCLASS	Table 6.3	Cross-tab ethnicity by social class, gender by table
		Table 6.4	
Gardiner and Hill (1996)	AREAP, CARS, ETHGROUP, LTILL, SEX, TENURE	Table 4	Logistic Regression car ownership by age, gender, house ownership, ethnicity and long-term limiting illness
Gardiner and Hill (1997)	AREAP, ETHGROUP, QUALEVEL, SEX, SOCLASS, TRANWORK	Table 1	Frequency table of cycling rate for area
		Table 2	Frequency table of cycling rate for area by gender
		Table 3	Frequency table of cycling rate for area by ethnicity
		Table 4	Frequency table of cycling rate by ethnicity, for Leicester
Gould and Jones (2000)	AGE, ETHGROUP, LTILL, RESIDSTA, SEX	Table 4 Model A	Multilevel logistic model for long-term limiting illness, level 2: area, by age, gender, ethnicity, and interaction terms
Green (1997)	DISTWORK, ETHGROUP, HOURS, INDUSDIV, OCCMAJOR, OCCSUBMJ, SEX, TRANWORK	Table 4.1	Frequency table of ethnicity by industry growth of employment, gender by table
		Table 4.2	
		Table 4.3	Frequency table of ethnicity by occupation growth
		Table 4.4	
		Table 4.5	Frequency table ethnicity by hours worked per week
		Table4.6	
Leventhal (1994)	AGE, CARS, CENHEAT, ECPOSFHP, ETHGROUP, RELAT, SEGROUP, SOCLASS, TENURE	Figure 1	Bar chart comparing prevalence of target group by region
		Figure 2	Marital status
		Figure 3	Ethnic group
		Figure 4	Social class
		Figure 5	Socio-economic group
		Figure 6	Availability of home central heating
		Figure 7	Number of cars

B Description of Subsetting and Sample Sizes used in the Papers

Paper	Table/Figure	Subset	Sample size-original
Ballard (1996)	Figure 5.1	Pakistani male	4,904
	Figure 5.2	Pakistani female	4,512
	Figure 5.3	Pakistani	9,416
	Figure 5.4	South Asian (Pakistani, Indian, and Bangladeshi)	29,724
Champion(1996)	Table 4.7	Migrants within Britain	95,177
Drinkwater and O'Leary (1997)	Table 5 males	Wales, age 16-64, resident of Britain, male	14,016
	Table 5 females	Wales, age 16-59, resident of Britain, female	9,843
Eade et al (1996)	Table 6.3	Age 16+, answered question on social class, not armed forces, male	340, 958
	Table 6.4	Age 16+, answered question on social class, not armed forces, female	292, 852
Gardiner and Hill (1996)	Table 4	Sheffield, Age 50+	3,532
Gardiner and Hill (1997)	Table 1	Answered question on travel to work and from selected areas	26,124
	Table 2		
	Table 3		
	Table 4	Answered question on travel to work, and from Leicester	2,048
Gould and Jones (2000)	Table 4 Model A	Resident of Britain, Aged 30-60	419,550
Green (1997)	Table 4.1	Male	540,967
	Table 4.3		
	Table 4.5		
	Table 4.2	Female	575,241
	Table 4.4		
	Table 4.6		
Leventhal (1994)	Figure 1	Age 16+	894,115
	Figure 2		
	Figure 3		
	Figure 6		
	Figure 4	Age 16+, answered social class question	647,323
	Figure 5		
	Figure 7	Age 16+, answered car question	866,097

C Synthesizing Order

AREAP, AGE, HOURS, COBIRTH, OCCSUBMJ, DISTMOVE, INDUSDIV, ETHGROUP, ECONPRIM, TENURE, OCCMAJOR, TRANWORK, DISTWORK, SOCLASS, RELAT, MSTATUS, WELSHLAN, CARS, RESIDSTA, QUALEVEL, QUALNUM, CESTSTAT, ECPOSFHP, SEGROUP, CENHEAT, SEX, DEPCHILD, LTILL

Given that REGIONP, DCOUNTY, and AREAP are hierarchical variables, REGIONP and DCOUNTY are not involved in the synthesis, since they can be derived from AREAP.

D Description of Variables from the Codebook

Variable name	Number of categories (excluding N/A)	Description
REGIONP	12	Individual SAR region
AREAP	278	Individual SAR area
ECONPRIM	10	Economic position (primary)
DISTMOVE	13	Distance of move-migrants
LTILL	2	Limiting long-term illness
TRANWORK	9	Mode of transport to work
SOCLASS	9	Social class (based on occupation)
CARS	3	Number of cars
OCCMAJOR	9	Occupation: SOC Major groups
OCCSUBMJ	22	Occupation: SOC Sub-major groups
INDUDSDIV	10	Industry (SIC Divisions)
COBIRTH	42	Country of birth
AGE	95	Age
MSTATUS	5	Marital status
SEX	2	Sex
DCOUNTY	63	Counties (Aggregations of SAR areas)
DISTWORK	7	Distance to work
ETHGROUP	10	Ethnic group
HOURS	72	Hours worked weekly
DEPCHILD	2	Number of resident dependent children
QUALEVEL	3	Level of highest qualification
WELSHLAN	5	Welsh language (Wales Only)
QUALNUM	3	No. of higher educational qualifications
TENURE	10	Tenure of household space
CESTTSTAT	3	Status in communal establishments
RESIDSTA	3	Resident status
RELAT	8	Relationship to household head
SEGROUP	20	Socio-economic group
CENHEAT	3	Availability of central heating
ECPOSFHP	3	Economic position of family head

E Single Imputation Totals

Paper	Table or Figure	syn 1	2	3	4	5	6	7	8	9	10	avg
Ballard (1996)	Figure 5.1	0.563	0.463	0.517	0.512	0.472	0.501	0.473	0.445	0.555	0.521	0.502
	Figure 5.2	0.544	0.538	0.525	0.562	0.5	0.509	0.524	0.54	0.528	0.536	0.531
	Figure 5.3	0.306	0.369	0.302	0.336	0.334	0.32	0.321	0.374	0.339	0.327	0.333
	Figure 5.4	0.253	0.266	0.24	0.245	0.289	0.263	0.237	0.284	0.28	0.293	0.265
Champion (1996)	Table 4.7	0.534	0.576	0.567	0.579	0.598	0.588	0.522	0.569	0.595	0.605	0.573
Drinkwater and O'Leary	Table 5 for males	0.511	0.501	0.518	0.562	0.542	0.495	0.523	0.536	0.506	0.542	0.524
	Table 5 for females	0.62	0.621	0.612	0.658	0.623	0.592	0.605	0.614	0.636	0.563	0.614
Eade et al (1996)	Table 6.3	0.334	0.321	0.361	0.35	0.332	0.348	0.369	0.344	0.337	0.334	0.343
	Table 6.4	0.365	0.368	0.346	0.349	0.345	0.371	0.341	0.362	0.351	0.344	0.354
Gardiner and Hill (1996)	Table 4	0.566	0.555	0.519	0.604	0.575	0.595	0.67	0.556	0.626	0.67	0.594
Gardiner and Hill(1997)	Table 1	0.381	0.319	0.457	0.336	0.277	0.345	0.388	0.348	0.313	0.294	0.346
	Table 2	0.516	0.497	0.554	0.493	0.467	0.52	0.519	0.513	0.477	0.526	0.508
	Table 3	0.53	0.453	0.566	0.48	0.508	0.451	0.408	0.48	0.483	0.467	0.482
	Table 4	0.53	0.635	0.599	0.714	0.698	0.648	0.563	0.516	0.538	0.549	0.599
Gould and Jones (2000)	Table 4 Model A	0.185	0.167	0.203	0.193	0.209	0.202	0.159	0.22	0.157	0.217	0.191
Green (1997)	Table 4.1	0.216	0.193	0.235	0.197	0.208	0.199	0.19	0.208	0.204	0.212	0.206
	Table 4.2	0.198	0.243	0.233	0.2	0.201	0.246	0.171	0.188	0.196	0.247	0.212
	Table 4.3	0.206	0.178	0.202	0.227	0.173	0.152	0.199	0.221	0.181	0.158	0.19
	Table 4.4	0.176	0.187	0.159	0.122	0.223	0.191	0.153	0.209	0.126	0.183	0.173
	Table 4.5	0.344	0.406	0.424	0.379	0.411	0.363	0.339	0.334	0.377	0.358	0.373
	Table 4.6	0.245	0.204	0.235	0.225	0.202	0.232	0.203	0.241	0.286	0.26	0.233

Table 8: Single Imputation CIO Scores

Paper	Table or Figure	syn 1	2	3	4	5	6	7	8	9	10	avg
Ballard (1996)	Figure 5.1	0.660	0.581	0.644	0.624	0.621	0.614	0.600	0.583	0.651	0.600	0.618
	Figure 5.2	0.564	0.602	0.604	0.600	0.560	0.602	0.595	0.594	0.605	0.609	0.594
	Figure 5.3	0.782	0.785	0.775	0.785	0.784	0.779	0.785	0.793	0.787	0.781	0.784
	Figure 5.4	0.819	0.816	0.807	0.811	0.830	0.815	0.813	0.817	0.821	0.835	0.818
Champion (1996)	Table 4.7	0.818	0.822	0.832	0.834	0.836	0.833	0.815	0.834	0.595	0.605	0.782
Eade et al (1996)	Table 6.3	0.787	0.782	0.799	0.790	0.783	0.787	0.796	0.791	0.791	0.786	0.789
	Table 6.4	0.779	0.782	0.781	0.775	0.777	0.786	0.774	0.784	0.785	0.776	0.780
Gardiner and Hill (1997)	Table 1	0.809	0.784	0.830	0.797	0.773	0.793	0.809	0.793	0.785	0.776	0.795
	Table 2	0.836	0.797	0.810	0.798	0.794	0.806	0.812	0.789	0.791	0.808	0.804
	Table 3	0.760	0.713	0.730	0.714	0.755	0.739	0.681	0.707	0.733	0.695	0.723
	Table 4	0.582	0.764	0.658	0.796	0.689	0.680	0.592	0.672	0.587	0.584	0.660
Green (1997)	Table 4.1	0.751	0.738	0.746	0.737	0.737	0.744	0.739	0.743	0.741	0.743	0.742
	Table 4.2	0.721	0.744	0.735	0.722	0.730	0.739	0.732	0.712	0.722	0.742	0.730
	Table 4.3	0.772	0.775	0.773	0.785	0.772	0.768	0.778	0.785	0.772	0.775	0.775
	Table 4.4	0.775	0.775	0.776	0.757	0.783	0.780	0.773	0.777	0.742	0.782	0.772
	Table 4.5	0.822	0.835	0.849	0.830	0.843	0.830	0.825	0.818	0.839	0.825	0.831
	Table 4.6	0.773	0.772	0.781	0.765	0.758	0.771	0.757	0.774	0.789	0.787	0.773
Leventhal (1994)	Figure 1	0.883	0.867	0.865	0.872	0.887	0.921	0.880	0.890	0.898	0.897	0.886
	Figure 2	0.780	0.778	0.742	0.776	0.740	0.736	0.785	0.773	0.749	0.738	0.760
	Figure 3	0.836	0.846	0.829	0.869	0.874	0.843	0.847	0.881	0.818	0.825	0.847
	Figure 4	0.734	0.712	0.728	0.702	0.758	0.708	0.715	0.731	0.738	0.756	0.728
	Figure 5	0.728	0.711	0.710	0.709	0.724	0.737	0.704	0.712	0.711	0.716	0.716
	Figure 6	0.883	0.872	0.876	0.876	0.885	0.909	0.879	0.879	0.921	0.918	0.890
	Figure 7	0.846	0.825	0.816	0.833	0.850	0.856	0.854	0.845	0.850	0.864	0.844

Table 9: Single Imputation ROE Scores