

LASI Synthetic Data Report

*Joshua Snoke, Erik Meijer, Drystan Phillips,
Jenny Wilkens, Jinkook Lee*

Paper No: 2023-009

CESR-SCHAEFFER WORKING PAPER SERIES

The Working Papers in this series have not undergone peer review or been edited by USC. The series is intended to make results of CESR and Schaeffer Center research widely available, in preliminary form, to encourage discussion and input from the research community before publication in a formal, peer-reviewed journal. CESR-Schaeffer working papers can be cited without permission of the author so long as the source is clearly referred to as a CESR-Schaeffer working paper.

LASI Synthetic Data Report

Working Draft

Joshua Snoke, Erik Meijer, Drystan Phillips, Jenny Wilkens, Jinkook Lee

September 2023

1. Introduction

The Gateway to Global Aging Data (Lee et al., 2021) is a data and information platform developed to facilitate longitudinal and cross-country studies on aging, especially those using the Health and Retirement Study and its International Network of Studies (HRS-INS) around the world. It is funded by the National Institutes on Aging (R01AG030153) and makes all of its resources available to researchers without cost.

The Gateway provides background information on the participating HRS-INS and each of their questionnaires. To enable comparisons across studies or longitudinally, the Gateway also provides concordance tables and comprehensive working papers detailing differences across studies and over time for similar measures. Beyond that, the Gateway develops and provides files that use a subset of the survey's individual-level data to produce Harmonized datasets, which contain variables defined to be as comparable as possible across surveys and over time. In addition to creating Harmonized datasets for the main or core interview, the Gateway has expanded into the creation of Harmonized datasets for life history interviews, harmonized cognitive assessment protocol (HCAP) interviews, COVID-19 interviews, and end-of-life interviews. For more detailed information, see Lee, Phillips, and Wilkens (2019).

A list of HRS-INS and some of their basic characteristics are given at the Gateway's website <https://g2aging.org/survey-overview>. The individual-level data of each study are available at no cost to researchers, after registration and signing a data user agreement. While the Gateway includes information *about* the studies (metadata), and prepares the Harmonized datasets and distributes the Stata code that generates these and their codebooks, users need to obtain the individual-level data from the original studies with few exceptions. One exception is the Longitudinal Aging Study in India (LASI), which began interviewing respondents aged 45 and older primarily from 2017-2019 with a subsequent interview planned to begin in 2023. As the Gateway does distribute the individual-level data for LASI, the barrier to access to the individual-level data is quite low. But for some purposes, such as use of the data in training workshops and webinars, it would be useful to have a version of the data with even fewer barriers.

In this document, we present such a dataset. This is a *synthetic dataset*, which means that none of the data correspond to real individuals. Instead the dataset is constructed purely from models that have been estimated on the original LASI data, such that the synthetic data resemble the original

data in terms of data structure (e.g., variable names) and statistical distributions of the variables. Thus, the synthetic data can be used to become familiar with the data structure, develop code, and explore relationships, without the hurdles that are involved in getting access to the original data. The synthetic dataset described in this document is modest in size, and expansions are planned, but it may already be useful for teaching purposes. The main goal of this document is to present these data as a proof of concept and to invite opinions from the research community about the usefulness of these synthetic data, and suggestions for improvements and extensions.

An important consideration is how precisely the distributions in the synthetic data correspond to the distributions in the original LASI data, and what the desired level of correspondence is in relation to potential disclosure of private information without the safeguards of the registration and data user agreement. If the models are very detailed, the synthetic data will precisely reproduce the original data, which would be undesirable. Conversely, if the models are too simple, distributions (esp. correlations and other forms of relations between variables) in the synthetic data may be very different from the distributions in the original LASI data, which would diminish the usefulness of the synthetic data.

In the remainder of this introduction, we briefly describe the LASI data, current data access regime, and potential confidentiality concerns with the synthetic data. Section 2 then discusses possible data access regimes for the synthetic data. Section 3 discusses the methodology for creating the synthetic data and section 4 contains empirical evaluations of the synthetic dataset, in relation to the distributions in the original LASI data. Section 5 concludes and discusses planned and potential next steps.

Background on LASI

The Longitudinal Aging Study in India (LASI) is a multidisciplinary survey of individuals age 45 and over and their spouses of any age in India (Bloom et al., 2021; Perianayagam et al., 2022). Wave 1 was mostly conducted between 2017 and 2019, and Wave 2 is in preparation as of May 2023. LASI was designed to be nationally representative, as well as representative of each of the 36 states and union territories in India. Therefore, the sampling was stratified by state and urbanicity (i.e., urban or rural) within state, and a three-stage clustered sampling design in rural areas and a four-stage clustered sampling design in urban areas. Sampling weights are provided that reflect the sampling design as well as differential response rates (although the latter are high compared to most other surveys, varying between 77% and 96% across states and territories). Once a household had been included into the sample, all household members age 45 and over and their spouses were asked to participate. Thus, the number of respondents per household varies across households, and is often higher than in comparable studies in the U.S. and other high-income countries.

LASI's questionnaire was designed to closely follow the U.S. Health and Retirement Study (HRS; Juster & Suzman, 1995; NIA, 2007), which started in 1992 as a nationally representative panel study of people over the age of 50 living in the United States and their spouses. Similar "international network of studies" of the HRS have been conducted in many countries across the world. The topics covered in the LASI questionnaire range widely and include demographics, health, socio-economic status, and social support, among others.

To further facilitate cross-country comparisons, the Gateway to Global Aging Data at the University of Southern California prepares harmonized datasets, with variable names and definitions, as well as codebook structure, closely following the RAND HRS (see Bugliari et al., 2023, for its latest version as of May 2023), which is a user-friendly data file that contains a large subset of the HRS variables. The starting point for the creation of our synthetic data is the Harmonized LASI data file (Chien et al., 2021).

Current Data Access Regime

The LASI data is currently accessible through the Gateway to Global Aging Data and through the International Institute for Population Sciences (IIPS) in India. Through the Gateway, the LASI data and Harmonized LASI can be obtained after website registration, completion of a data use agreement, and brief confirmation of identity. Through IIPS, the LASI data can be obtained after completion of a Data Request Form, including a study proposal for use of the data and valid proof of identity, and its subsequent approval.

While nearly all data is publicly released, district-level (referred to as SSU) information would increase the probability of respondent identification and so has been restricted. This information is not currently available outside the LASI study team, but may become available upon the establishment of the Gateway to Global Aging Data Enclave, which is currently in progress.

Privacy and Confidentiality Concerns

As part of collecting the LASI, respondents are ensured that their privacy and confidentiality will be maintained throughout any downstream publications or data sharing. This is vital both to preserve the ethical right to privacy and also to ensure the viability of future data collections. Respondents might be concerned about re-identification, such that someone could ascertain that they participated in the survey. Along with this, they may be concerned about disclosure of certain sensitive attributes which may result from re-identification. In other cases, attributes may be learned by inference, such as the specific income of the wealthiest person in a small town.

With a very large underlying population, such as India's, it may be tempting to think that the risk of such disclosures is minimal. But given enough information, particularly geographically identifying information in low population regions, studies have shown that disclosure is quite possible (e.g., Rocher et al., 2019). Additionally, the LASI relied on substantial oversampling of places with small populations in order to allow geographically representative analyses. While in some large states, the sampling fraction is very small, in smaller states it is larger. Additionally, even with small sampling fractions very unusual records can lead to disclosures. For these reasons, we cannot simply assume releasing synthetic data removes all risk for participants, and we conduct disclosure risk assessments as part of the development of a synthetic data file.

2. Possible Data Access Regime(s)

Using synthetic data opens the door to a few potential means for increasing data access. One possibility is to maintain the existing registration process but to include additional variables in

the synthetic data that are not currently part of the harmonized LASI. This may include LASI variables or data merged from other sources, which is currently too sensitive to share without taking additional protection measures. The synthetic data in this application could comprise only the sensitive variables, known as partially synthetic data, which would be matched to the regular data by their unique identifier, or the LASI variables could be synthesized along with the merged variables. This approach allows registered users to explore relations and develop code without additional registration requirements without unacceptably increasing disclosure risk. In many cases, once the user has finished their code, they will run the code on the confidential data in a more controlled environment.

Alternatively, synthetic data could be released as a fully public dataset which is downloadable online without any restrictions. In this case, the privacy protections from the synthetic data models themselves should be stronger, and it may not be possible to include as many variables or granularity of information as in the first framework. In this application, all the variables in the file will be fully synthetic. The benefit of this approach is that it increases overall data access by lowering the barriers, and it allows researchers to freely share the synthetic data with collaborators. Moreover, this approach allows educators to use the data easily in their teaching and share (processed versions of) the data with students.

Compared to the second approach, the advantage of the first approach is that the existing access framework can be utilized. The registration requirement then offers additional protection against misuse of the data by narrowing the pool of users and incorporating data use agreements which can be used to enforce penalties.

This report provides a proof of concept that describes a fully synthetic dataset of variables that are in the regular data, and we leave decisions about the access model for later work. These decisions should be based on the feasibility of sufficiently protecting the confidentiality of survey respondents as well as the use cases for the synthetic data.

Use Cases

While there are many potential uses for synthetic data, we have envisioned three major avenues for the use of synthetic Harmonized LASI data at this time.

- (A) User trainings: The Gateway conducts quarterly user trainings to teach students or researchers how to utilize Harmonized data for their analyses. Rather than requiring participants to pre-register for a study, which is usually met with limited success, the utilization of synthetic data would limit the need for pre-registration. The synthetic data used for this purpose would contain a limited number of variables, just sufficient enough to complete several sample analyses and to give participants a sense of the data structure. Once they have gone through the training, they can apply for the Harmonized LASI or any other Harmonized data with a better understanding of the data and how to conduct their research.
- (B) Exploratory research: We could make the synthetic data easily available on the Gateway website. In that way, researchers could conduct exploratory research and determine

whether the data would suit their research plan. If they find it suitable, they could then apply for the Harmonized LASI and use it in place of the synthetic data for their analyses.

- (C) Restricted data research: We could use synthetic data to protect respondent anonymity when using restricted data. While the Harmonized LASI releases data at the state-level, many potential analyses require more specific geographic locations, especially when linking to the LASI community survey. Additionally, data on pollution or severe weather are very closely tied to specific geographic locations, potentially making respondents identifiable, especially when historic records and residential history are available. The use of synthetic data in these cases would allow researchers to conduct their analyses in full detail, while protecting respondent anonymity. This type of use would require registration and data use approval.

3. Proposed Data Synthesis Methodology

In this chapter we provide an overview of the methodology we use for generating a synthetic version of a subset of the LASI. We give a brief background on general synthetic data concepts, discuss the data selected for the initial demonstration product, detail our approach to maintaining the multiple levels of observation (households, couples, individuals) in the data, and finally describe the specific models we use to generate synthetic values.

Background on Synthetic Data

In order to increase data access, it is common for data maintainers to apply statistical disclosure control (SDC) methods to confidential data. Approaches include data reduction methods such as suppressing certain variables, records, or cells (for tabular data) or coarsening variables by collapsing categories. Other methods perturb records to protect privacy, such as data swapping which randomly switches values between records while preserving the marginal distributions, bottom- or top-coding that replaces outliers with selected minimum or maximum values, and random noise addition that adds noise to variables from known distributions. See Hundepool et al., 2012 for a comprehensive review of these methods. These methods attempt to reduce the risk to breaches of confidentiality in order to share versions of microdata, tables, or summary statistics that otherwise could only be accessed through restricted means.

Synthetic data was proposed as an alternative method for allowing researchers to access microdata while minimizing the risk of disclosure from releasing that data (see Rubin 1993, Little 1993, Raghunathan et al., 2003). Conceptually it differs from other SDC approaches. Rather than starting from the entire confidential sample and attempting to maintain as much of the original data values as possible, synthetic data starts with sample parameters based on the data, according to some assumed data generating process, and draws fully new records based on a model using these sample parameters. In this way, a model and a data generating process are essential to synthetic data in a way that they are not for other statistical disclosure control methods.

Synthetic data shares some similarities to missing data imputation and micro-simulation, but it differs both in terms of what data are used to create the models and the target of the models. In synthetic data, the complete observed data is often available (in some case imputation and synthesis are combined sequentially), and the goal is to produce a data set that captures the same

statistical properties of the observed data rather than filling in missing information or simulating future information.

The original methods proposed releasing synthetic data in place of the confidential data, and they provided rules for valid inference such that analysts could fit valid models without ever accessing the confidential data. But inferences from synthetic data are only valid if the models that are used to synthesize the data correspond to the true underlying data generating process (Raab et al., 2016). Due to the difficulty of validating this assumption, synthetic data are more commonly used as exploratory data sets. After testing models using the synthetic data, researchers will confirm results using restricted access data or verification servers (Reiter et al., 2009). We do not cover valid inference methods in this report, but Drechsler (2011) or Raab et al., (2016) provide extended details on this topic.

Assuming the synthetic data will be used primarily to explore relationships in the data, the synthetic data model must either attempt to forecast the analytical models of interest to researchers in order to accurately include them in the model, or it must attempt to generally capture the overall joint distribution (see Snoke et al., 2018). For the purposes of this proof-of-concept LASI synthetic data file, we are more concerned with capturing the general distributions of variables in the data. Future work that includes additional variables may focus more on capturing specific relationships in the data.

We use the common approach of fully conditional sequential (FCS) models for synthesis, which means that we model and generate new values for one variable at a time in a sequence that approximates the joint distribution of the variables. This is based on the law of total probability:

$$P(X, Y, Z) = P(X) P(Y|X) P(Z|X, Y)$$

This gives us flexibility to choose separate models for each variable, and it is a common approach for similar approaches to missing data imputation. For example, consider the data set containing the variables:

$$\{State, Gender, Income\}$$

Leaving *State* un-synthesized (i.e., the number of records per state is the same as in the original data), one could first fit a logistic regression model for the binary outcome *Gender* using *State* as the predictor. New values of *Gender* are then sampled using the fitted coefficients to produce synthetic gender values that capture the distribution across states. Next, one would fit a (log) linear regression model for the outcome *Income* using *State* and *Gender* as predictors. Synthetic values of income are drawn using the fitted coefficients, as well as the previous drawn synthetic values of *Gender*. These models approximate the distribution.

$$f(Gender, Income|State) = f(Gender|State)f(Income|Gender, State)$$

As seen in this example, one may choose whether to synthesize all variables that comprise a record, generally known as *fully* synthetic data, or to include some confidential variables without synthesis, known as *partially*, synthetic data. It is also possible to include additional variables in

the synthesis models which are neither synthesized nor released in the synthetic data. These are known as *auxiliary* variables. For example, if we chose not to release *State* in our above example, it would be an auxiliary variable. For our current demonstration product, we create partially synthetic data, leaving *State and Rural indicator* un-synthesized. We also utilize one auxiliary variable. The following section provides specifics on the variables we use.

Candidate LASI Data for Synthesis

We synthesize the following variables in this demonstration version. Future work will include additional variables. The file contains both household and individual level characteristics, as well as couple identifiers indicating spousal respondents. There is one auxiliary variable, *Head of HH gender*, which was not part of the original Harmonized LASI file, but was derived to be used in the synthesis process to reweight the data after synthesis to match the original weight totals. It will not be released in the synthetic data.

<i>Variable type</i>	<i>Variable</i>
Household Level Variables	State
	Rural indicator
	HH total size
	HH # respondents
	HH survey weight
Individual Level Variables	Gender
	Age
	Education years
	Education category
	Self-reported health
	Activities of daily living
	Working status
	Individual earnings
	Couple ID
	Individual survey weight
Auxiliary Variables (Not Released)	Head of HH gender

Table 1. Variables Included in the Synthesis

Maintaining HH and Couple Structure

The LASI file has a household structure with couples and individuals within the households. While we will release an individual-level file, crucially we want to maintain both the household and couple structures in the synthetic data. To do this we use a novel approach to synthesizing the structure by reshaping the data in particular ways and add some structural variables to the synthesis process. Prior work on synthesizing household structure is limited. Hu et al., (2018) proposed using hierarchical Bayesian models which are computationally intensive and require defining complex joint distributions for the variables in the data. Benedetto and Totty, (2020) presented a method where couples are matched after synthesizing based on a clustering algorithm. The structure in our data set, which contains both households and couples, is more

complex than in either of those prior works, and we do not see a straightforward means of applying those methods to our data.

Instead, we take a simpler but intuitive approach to maintaining the structure in the data. First, we restrict the total number of individuals that can exist in the survey from each household. For example, if we restrict to at most three respondents this captures 99% of the individuals in the original data because very few households contain more than three individuals in the survey.

In order to allow for different couplings of the maximum allowable individuals, we synthesize variables for each possible individual “slot”. With a maximum of three respondents, the five possible individual slots are:

1. First respondent, first couple
2. Second respondent, first couple
3. First respondent, second couple
4. Second respondent, second couple
5. First respondent, third couple

For households of size 1, only the first individual is ever synthesized. For households of size 2, the only possible combinations of individuals are slots (1, 2) and (1, 3). For households of size 3, the possible combinations of individuals are slots (1, 2, 3), (1, 3, 4), (1, 3, 5). Table 2 illustrates the possible person combinations that exist with a maximum of three respondents per household.

HH Respondents	Person 1 Age	Person 2 Age	Person 3 Age	Person 4 Age	Person 5 Age
1	50	NA	NA	NA	NA
2	45	47	NA	NA	NA
2	45	NA	65	NA	NA
3	47	45	66	NA	NA
3	70	NA	48	46	NA
3	70	NA	65	NA	47

Table 2. Possible HH Person Patterns with Age as an Example Variable

Using this coding, we synthesize entire household rows rather than individuals within households. As described in more detail in the next section, we use FCS synthesis models as described earlier, such that we synthesize one variable at a time with prior variables as predictors. By using this approach, we capture the relationships between individuals in the same households and individuals who are couples.

To ensure values exist that are consistent with the corresponding synthesized respondents in the household and the possible combinations we first synthesize “existence” variables based on the household characteristics, i.e., binary variables for “person slot 1 exists”, “person slot 2 exists”, etc. We can easily extend this to any number of person slots depending on how many respondents we allow per household. We then fit synthetic models for individual-level variables only on the corresponding subset of households which contain the corresponding person slot. For example, to synthesize individual characteristics for “person slot 2”, we first subset to the

confidential households where person slot 2 exists to fit the models. We then only synthesize new values for the synthetic households where person slot 2 exists. This enforces structural missingness such that these existence variables match the total synthesized number of survey respondents in the household and one of the possible combinations shown above.

A potential drawback of this approach is that we must choose a fixed number of possible household respondents. In practice we find it makes sense to set this lower than the observed maximum number of household respondents in the data because the number of household respondents has very small outliers. Given that such a small number of households contain more respondents, it would be difficult to model them accurately and if we did it would present a privacy risk. Accordingly, we do not see setting a smaller upper limit on household size as a significant drawback. We present evaluations in Section 4 on the potential impact on utility and risk.

Synthesis Model and Predictor Selection

Currently we use flexible non-parametric synthesis models using the CART algorithm first proposed for synthesis by Reiter, (2005) and implemented in the *synthpop* package in R (Nowok et al. 2016a; Nowok et al. 2016b). All released variables are synthesized except for *State* and *Rural indicator*. By not synthesizing state and rural/urban indicators, we maintain the same number of households by state and rural/urban areas in the synthetic data as are in the original data. These two geographic variables formed the core of the sampling frame of the confidential survey data.

The predictors for each variable are shown in the appendix. Rather than the usual approach of using every preceding variable in the synthesis sequence as a predictor, we simplify the predictor set slightly. Household variables are synthesized first (other than state and rural/urban), with each preceding HH variable as a predictor. Next, person existence variables are synthesized based on HH variables and the other existence variables.

Individual variables are synthesized conditional on (1) all household variables except the hh weight, (2) all preceding individual variables for a given individual, and (3) all matching individual variables for other preceding individuals (e.g., age of individual 2 is conditional on age of individual 1). Due to the structure of the persons described earlier, individual 4 and 5 variables are conditional on individual 1 and 3 only because households with individuals 2 and individual 4 or 5 are structurally impossible. Finally, individual weights are synthesized conditional on HH weights rather than the other HH variables. Future work can explore the impact of these predictor choices.

When synthesizing the individual characteristics, we also make additional subsets to the data in order to ensure the structure of the data. First, when fitting the prediction models for individual slots 2 through 5, we only estimate them using the confidential records where the predicted person slot exists. Second when predicting new values, we only make predictions where the person existence variables were synthesized to be true.

After generating the synthetic data, we perform a final reweighting of the individuals and households. We do this so that the synthetic data have the same representativeness of the

confidential data. Detailed information on the generation of LASI individual and household weights can be found in Chien et al., (2021). We use the same raking and trimming approach, and we use the household and individual weights generated from the synthesis process as our starting weights in order to capture the original variation in the confidential starting base weights.

The household weights are calibrated to the household population for one factor within each state:

- *Head of household gender x Rural indicator*

The individual weights are calibrated to the individual population for three factors:

- *Gender x Age*
- *Rural indicator*
- *Gender x Education*

The resulting weights provide representativeness on the weighted characteristics both within states and nationally across all states. We use the weight distributions in the confidential data to estimate the population distributions across the raking factors.

Additional Synthesis Tuning Parameters

In addition to the general synthesis framework we described, we made a few additional practical decisions when it came to synthesizing the data. Some of these were made for computational purposes while others are possible parameters to control the trade-off of utility and risk.

First, due to the large number of unique states in the data, we stratified our synthesis into four subgroups of states based on regions (North, Central, East, South). This significantly increased computational speeds, and due to the inclusion of state indicators in our predictors it should not have adverse effects on the accuracy of the models. In other words, we still model state-level variation within each region. The potential drawback is that by fitting smaller models, we lose some potential privacy protection from smoothing over a large set of households across a larger geographic area. We may also lose statistical precision for modeling the population when using subsets, but each region still represents a large, diverse sample of states and households.

Second, we binned age into 5-year bins (with open-ended bins below 45 and above 80). We synthesized the bins and then drew random values from within the bins afterwards based on the empirical distribution. We made this step as a type of smoothing, and we plan to explore more formal means of smoothing over ages.

Lastly, we tested a few different values for the CART model minimum bucket size parameter. This controls how many observations each terminal node must have in order to make an additional split. We use a minimum bucket size of 15 for our first demonstration product.

4. Evaluations

We first show evaluations of one synthetic data version. Later in this section we show comparisons when further changes are made to the generation process that provide different risk-utility tradeoffs.

We measure the utility of the data using distributional comparisons of individual, spousal, and household characteristics. We primarily focus on univariate and bivariate comparisons of the variables in the data, but we also compare the results of specific analyses one might run using the synthetic data. We compare both weighted and unweighted results.

Single Variable Distributional Comparisons

We start by comparing the univariate distributions for each of the variables in the data. Table 2 provides some selected summary statistics. Histograms comparing the confidential and synthetic data can be found in the Appendix in Figures A.1-A.3. We do not see any significant differences in the distributions.

<i>Variable</i>	<i>Confidential</i>	<i>Synthetic</i>
Gender		
% <i>female</i>	57.8 [57.4, 58.1]	57.6 [57.3, 58.0]
Education category		
% <i>Less than lower secondary</i>	70.4 [70.1, 70.8]	70.7 [70.4, 71.1]
% <i>Upper secondary/vocational</i>	24.4 [24.0, 24.7]	24.1 [23.8, 24.5]
% <i>Tertiary</i>	5.22 [5.06, 5.39]	5.11 [4.95, 5.28]
Working status		
% <i>working</i>	43.4 [43.0, 43.8]	43.1 [42.8, 43.5]
HH # residents	4.80 [4.77, 4.82]	4.79 [4.76, 4.81]
HH # respondents	1.67 [1.66, 1.67]	1.66 [1.66, 1.67]
Mean age	57.8 [57.7, 57.9]	57.9 [57.8, 58.0]
Mean education years	4.32 [4.29, 4.36]	4.28 [4.25, 4.32]
Mean self-reported health	3.19 [3.18, 3.20]	3.19 [3.18, 3.20]
Mean activities of daily living	1.26 [1.26, 1.27]	1.26 [1.26, 1.27]
Mean income	\$27,800 [\$26,600, \$29,000]	\$28,000 [\$26,800, \$29,200]

Table 2. Selected Summary Statistics for Confidential and Synthetic Data

We can summarize the distributional closeness between the confidential and synthetic data using the standardized *propensity score mean squared error* (pMSE) (Snoke et al., 2018) metric. Details on the computation of this metric can be found in the Appendix. A value of 1 indicates that the distribution of synthetic values are the same distance from the confidential data as would be expected on average for new samples drawn from the same distribution that generated the confidential data. While no upper bound value exists, Raab et al., (2021) suggest that values below 10 provide good utility as a rule of thumb.

Figure 4 shows the standardized pMSE computed over the univariate distribution for each variable. This allows us to see which variables are distributionally closer to the confidential data, taking into account the variation in the underlying data generating process.

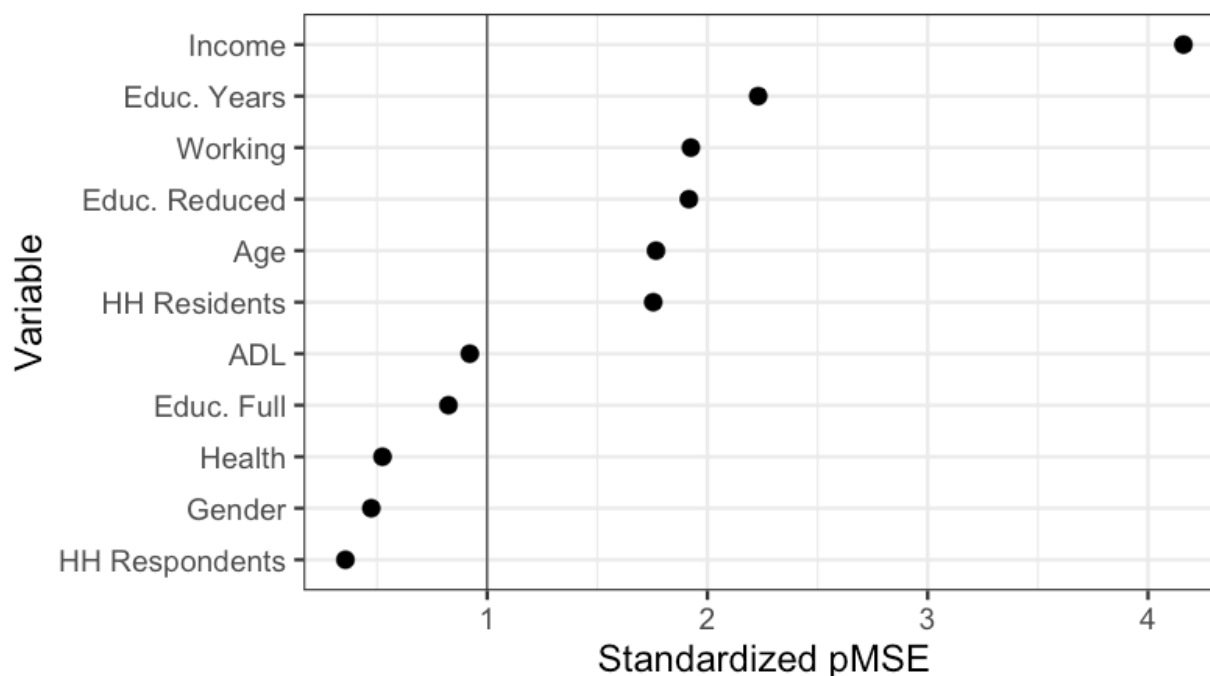


Figure 1. All univariate standardized pMSE values for individual or household variables in the synthetic data.

We see that many variables have values close to 1. The largest value is for *Income* followed by *Education Years* (continuous). These results alone do not suggest any improvements need to be made, but more detailed evaluations are important. In particular, univariate evaluations do not necessarily tell us how well the data will replicate common use cases of the data.

Couple and Household Distributions

We evaluate how well the synthetic data captures the couple and household distributions in the confidential data by looking at the relationships between the male and female partner in the couple¹ and different persons in the household. First, Figure 2 shows the ages of the female and male partners in the couple. Apart from a few individuals where the female partner is significantly older than the male, the synthetic data accurately captures distribution in the confidential data.

¹ There are a small number of same-sex couples in the data, but we do not report results on these individuals due to confidentiality concerns. The synthetic data captures approximately the same rate of same-sex couples observed in the confidential data.

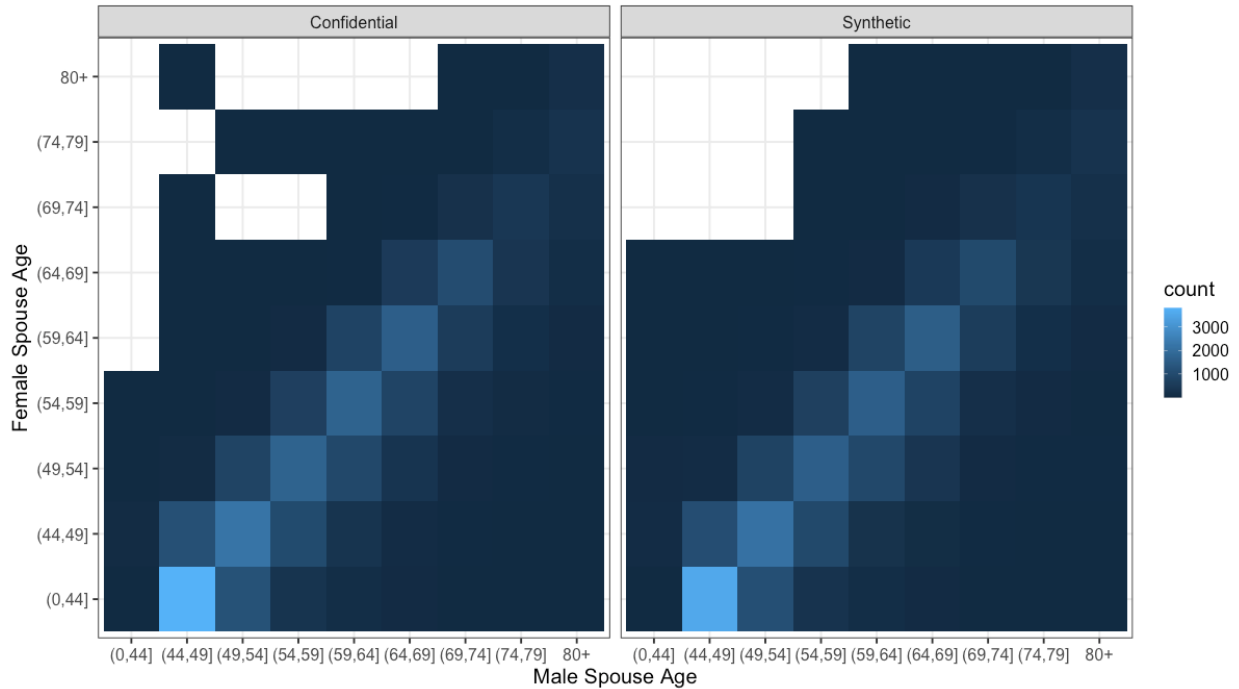


Figure 2. Distribution of the ages of the female and male partners in the couple in the confidential and synthetic data.

Next, Figure 3 shows the difference between the educational attainment for the male and female partners in each couple. Again, the synthetic data captures the distribution accurately. It slightly underpredicts the number of couples with the same number of years of education, and the errors appear roughly even distributed on either side of zero.

Switching to household comparisons, Figure 4 shows the bivariate distribution between the oldest and youngest household member, in households with more than 1 member. The synthetic data does an even better job, relative to the couples, of preserving the household relationships.

Finally, Figure 5 shows the difference between the most and least years of education of household members. We see that similarly to Figure 3, the synthetic data slightly underpredicts the number of households where the difference is zero, but overall the synthetic data does a very accurate job of replicating the distribution of educational attainment for individuals within the same household.

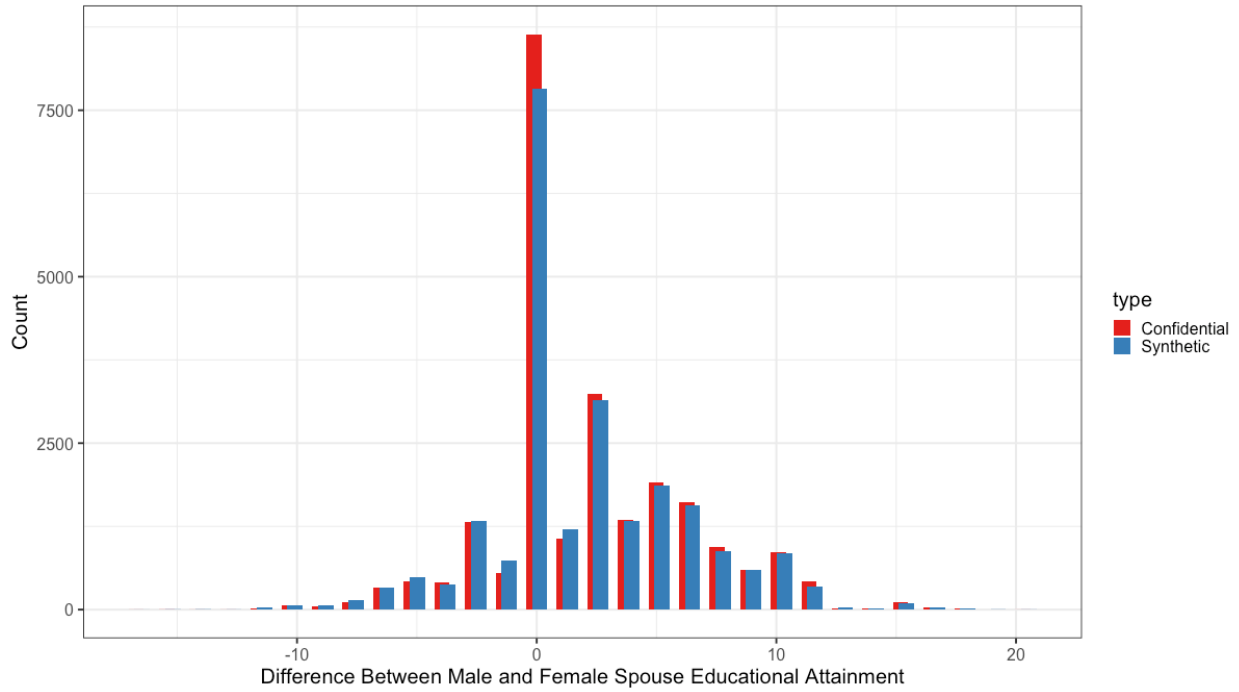


Figure 3. Distribution of the educational attainment (in years) of the female and male partners in the couple in the confidential and synthetic data.

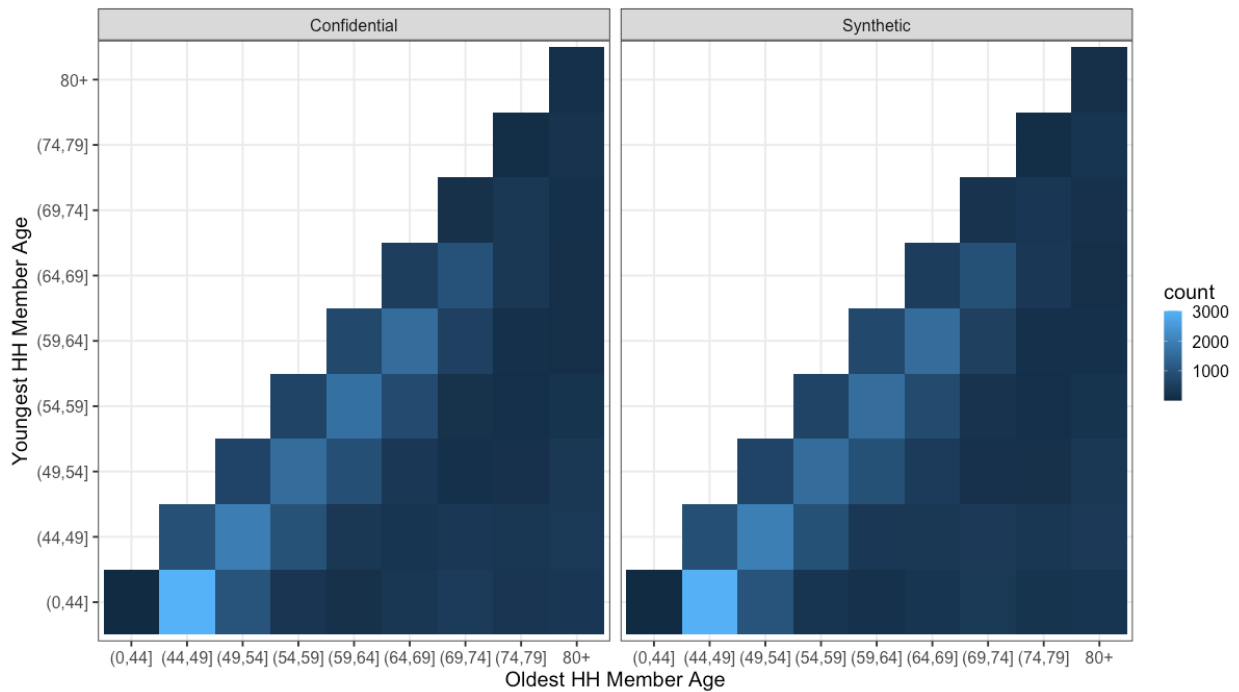


Figure 4. Distribution of the ages of the oldest and youngest individuals in the household in the confidential and synthetic data.

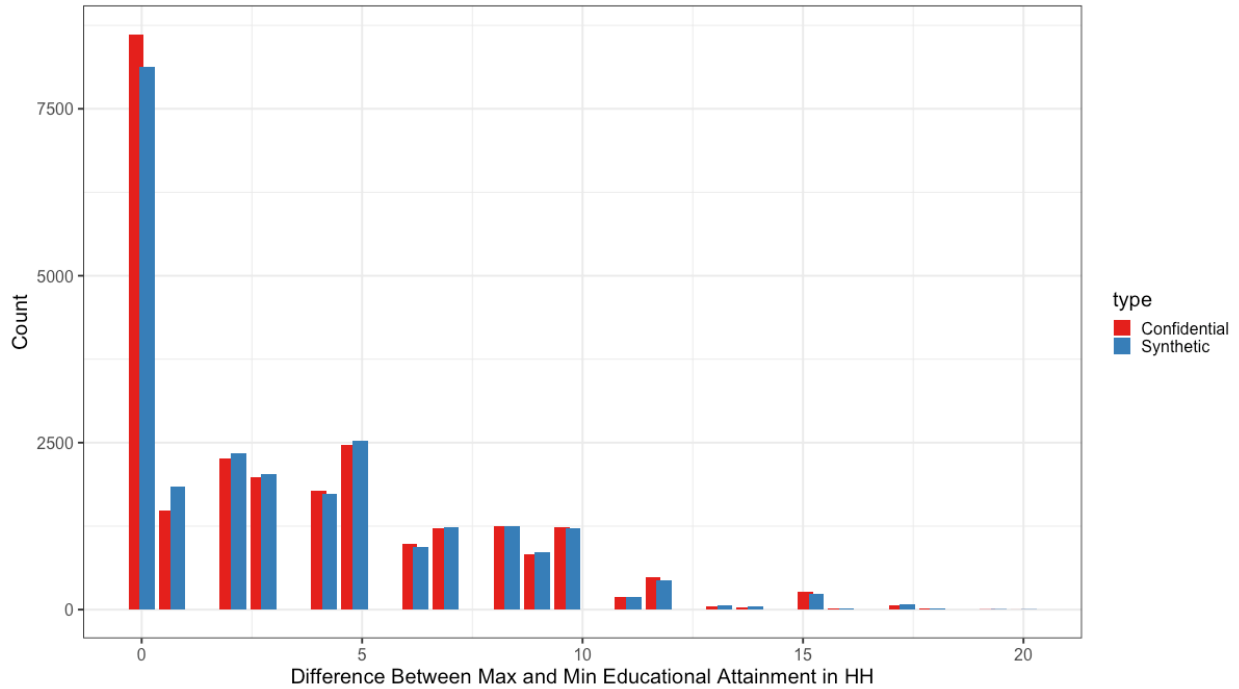


Figure 5. Distribution of the educational attainment (in years) of the individuals in the household with the most and least educational attainment in the confidential and synthetic data.

Bivariate Distributional Comparisons

Next, we consider the similarity of bivariate distributions in the synthetic and confidential data. This helps us understand how well the synthetic data model captures correlations between pairs of variables. We summarize results using the same standardized pMSE distance metric, where 1 is the target value. This metric can be used across both continuous and categorical or binary variables, which makes it more applicable than comparing correlations.

Figure 6 shows the results for all two-way interactions. We see now that while the marginal distribution of the number of ADLs was captured well, the bivariate relationships with this variable are the least accurate in the synthetic data. Analyses using ADL are likely to be less similar to the confidential data than analyses using other variables in the data. As a variable of interest for many analyses, additional work is likely needed to improve the synthetic data models generation of ADL values.

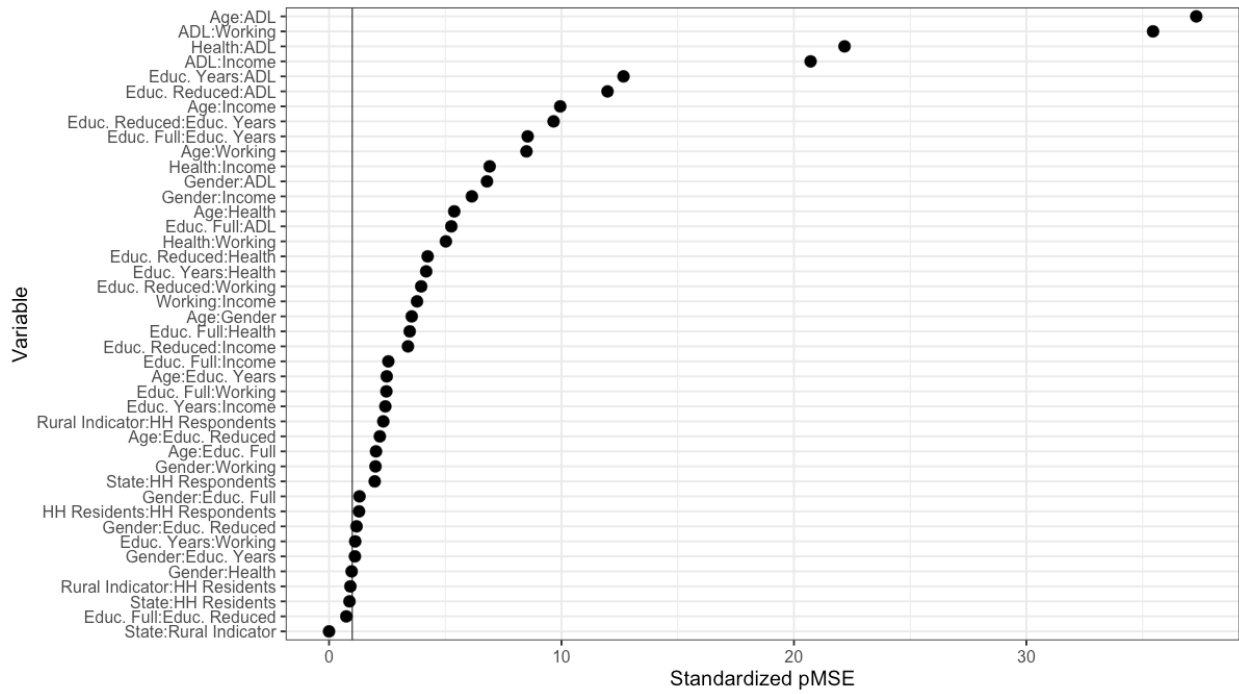


Figure 6. All bivariate standardized pMSE values for individual or household variables in the synthetic data.

Figure 7 shows the average value for each variable across the bivariate relationships (shown in Figure 6) that include the variable. We see that on average, the distribution of *ADL* with other variables is the worst captured in the synthetic data. These results suggest that the synthetic data models could be improved when it comes to capturing the distribution of *ADL* conditional on other characteristics. *Age* and *Working status* also have higher discrepancy between the synthetic and confidential data.

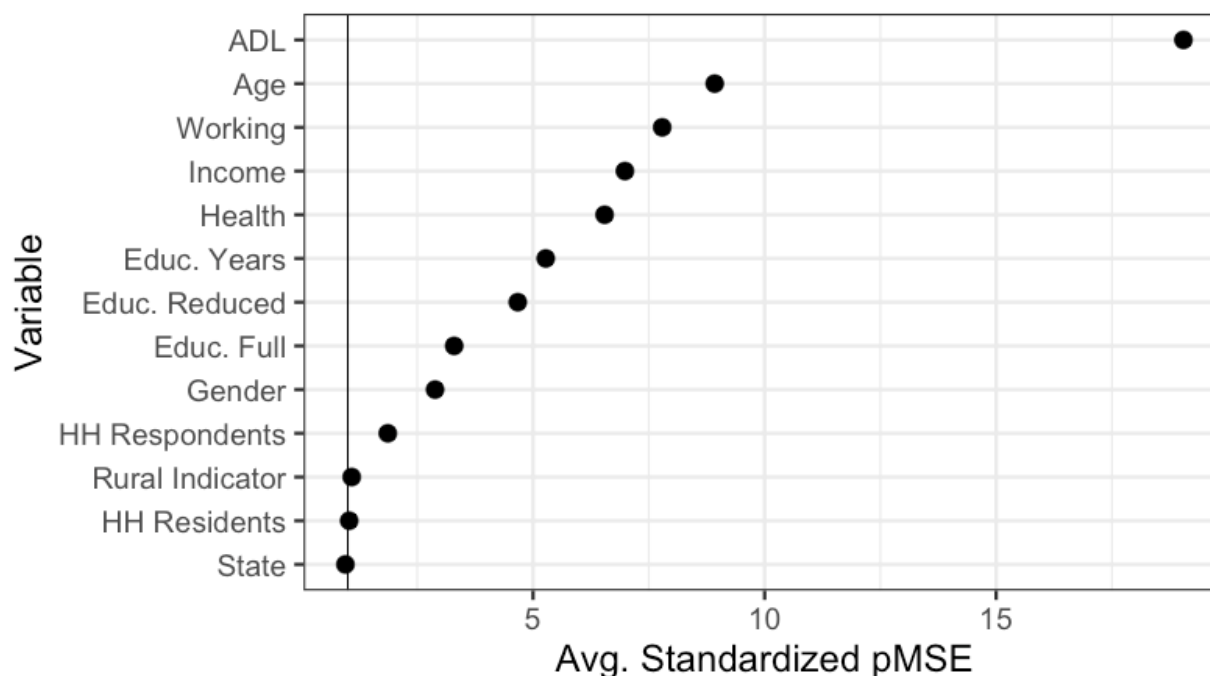


Figure 7. Average values across all bivariate standardized pMSE values for each individual or household variables in the synthetic data.

Three-way Distributional Comparisons

Evaluations comparing distributional similarity of all three-way interactions are shown in Figure A.5 in the Appendix. They show a similar pattern as the bivariate results for the average value across each variable. We see that in particular, the distribution of ADL with working status, gender, and age are poor relative to the other three-way distributions.

Model Specific Results Comparisons

Finally, we inspect a specific model to get a better idea of how the synthetic data replicates (or does not replicate) the results in the confidential data. We fit a simple model predicting ADL using gender and age as covariates, which is a common model used in LASI trainings. Given the bivariate and three-way distributional results shown in the previous sections, we expect this model will not give as accurate results as we might expect if we used other variables in the synthetic data. In fact, as we see in Table 3 and Figure 8, the synthetic point estimates move towards zero, and the confidence intervals do not overlap with the confidential estimates.

Coefficient	Confidential	Synthetic	Relative Difference
Intercept	-0.8384	-0.2414	-71%
Age	0.0179 [0.0174, 0.0184]	0.0084 [0.0079, 0.0089]	-53%
Gender	0.1186 [0.1064, 0.1308]	0.0321 [0.0199, 0.0443]	-73%

Table 3. Coefficients for the ADL model estimated on the confidential and synthetic data with relative differences between the coefficients.

Although the Z-scores are smaller, shown in Appendix Figure A.6, the sign and significance do not change, so a researcher using the synthetic data would still find the same type of effect. But the estimated effect size is significantly smaller in the synthetic data. Using the relative difference, the estimated effects of age and gender are between 53 and 73% smaller. This is due to the synthetic data models shrinking the correlation that exists between these characteristics in the confidential data.

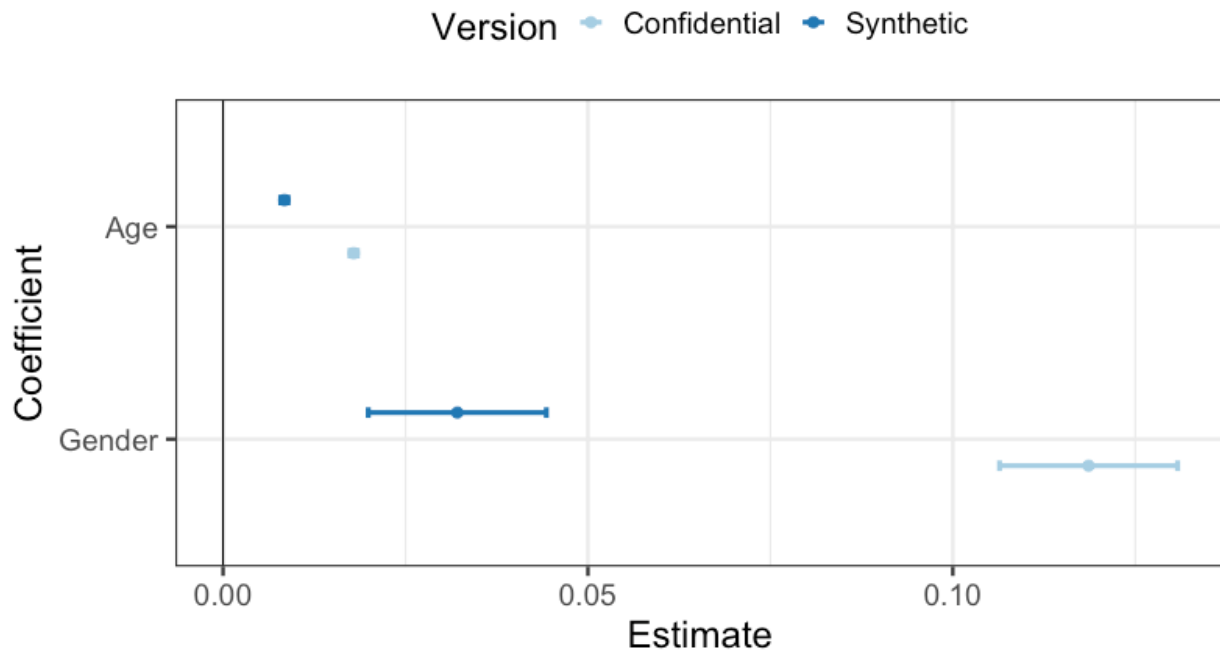


Figure 8. Depiction of the point estimates and confidence intervals for the two predictors in the ADL model, estimated using the confidential and synthetic data.

Additional Model Tuning Parameters for States with Small Populations

A few states and territories in the data have much lower populations, but they were significantly oversampled in order to create a representative sample of every Indian state and territory. For these locations, the disclosure risk from replicating similar synthetic data records to the confidential records is higher because there are fewer other potential candidate households from these states. We consider two approaches to mitigating this risk.

First, we consider grouping small states with larger neighboring states. This straightforward approach removes any unique relationships in the data that exist within the smaller states and instead draws values based on the average relationship between the smaller and larger states. This decreases the expected utility of both the smaller and larger states.

The second approach uses the sampling rates to down-weight records in the synthesis process for households that were sampled with high probability. (Hu et al. 2021) used a similar approach that down-weighted “risky” records based on their contribution to the likelihood in a Bayesian model. We use the inverse of the base sample weights (different from the weights included in the data file) as case-weights in the CART models. We cap the maximum value at 100 and set any

values below 10 equal to zero. This removes any contribution of records sampled with probability greater than 0.1 from the synthesis model. Because the models are estimated over groups of states, down-weighting records will have the effect of capturing relationships in small states that are similar to those in larger states but reducing the ability of the model to learn unique relationships in smaller states and reflect these in the synthetic data. It should not have a substantial impact on larger states.

Figures 9-11 show results for two states, one smaller and one larger, that were grouped together in the grouped scenario. In each figure the left panel shows the smaller state utility and the right panel shows the larger state utility. Each show both the results for (1) the full state synthesis (same as shown above), (2) the grouped synthesis, and (3) the down-weighted synthesis.

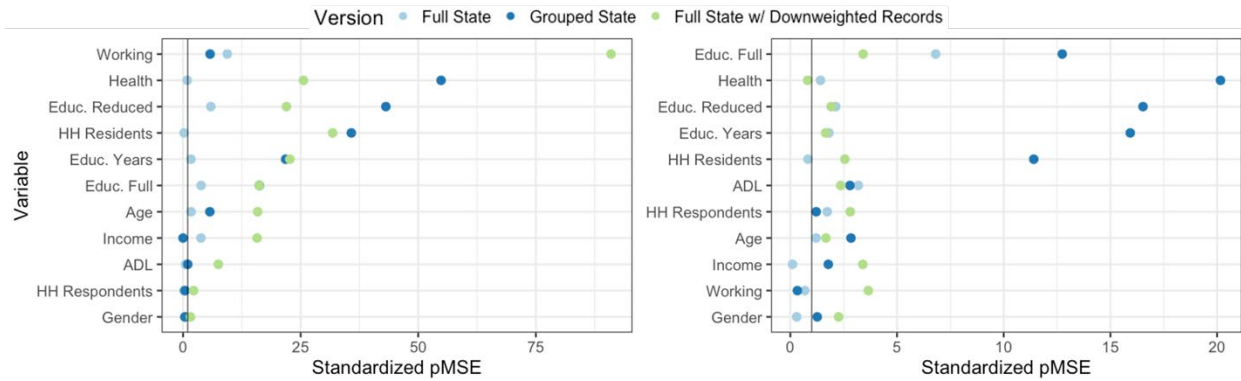


Figure 9. All univariate standardized pMSE values for individual or household variables in the synthetic data. Showing synthesis models using all states, grouped states, and down-weighted records. Left panel showing a small state and the right panel showing a large state, which were grouped together in the grouped synthesis.

We see as expected that both grouping and down-weighting reduce the utility (higher standardized pMSE scores) for the smaller state. The down-weighting approach performs better for most variables, though it is quite poor for *Working status*. Also as expected, the utility only gets worse for the grouped approach for the larger state, as shown in the right panel. This suggests that the down-weighting approach may be a more targeted means of protecting small states.

Figure 10 shows the average standardized pMSE across each of the three-way interactions. Overall the down-weighting approach has better utility in both the smaller and larger state. In particular, the household level variables' utility is quite poor for the grouped approach.

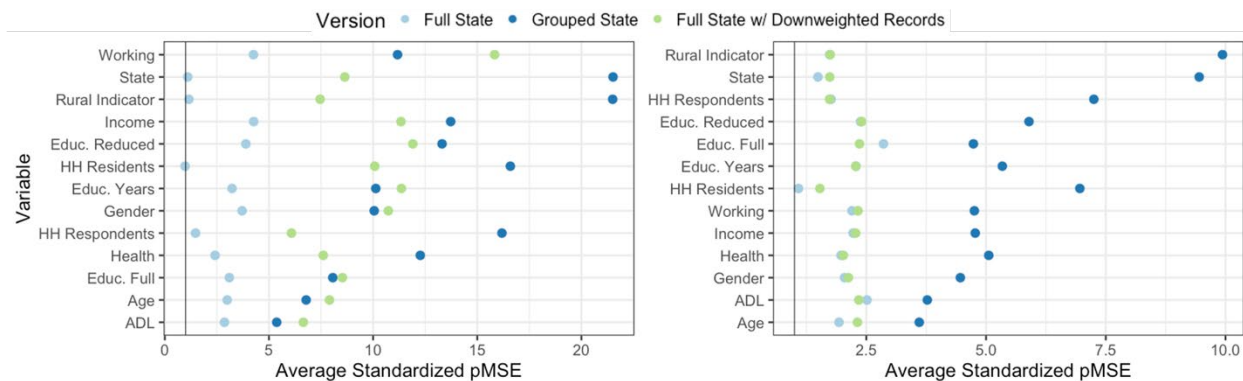


Figure 10. Average values across all three-way standardized pMSE values for individual or household variables in the synthetic data. Showing synthesis models using all states, grouped states, and down-weighted records. Left panel showing a small state and the right panel showing a large state, which were grouped together in the grouped synthesis.

Impact of Re-Weighting Post-Synthesis

Finally, we consider the impact of re-weighting the records to match known population targets after the synthesis is done. As a reminder, we re-weighted to match the targets which the original data were weighted to meet, which included (1) *Head of household gender x Rural indicator* for households and (1) *Gender x Age*, (2) *Rural indicator*, and (3) *Gender x Education* for individuals.

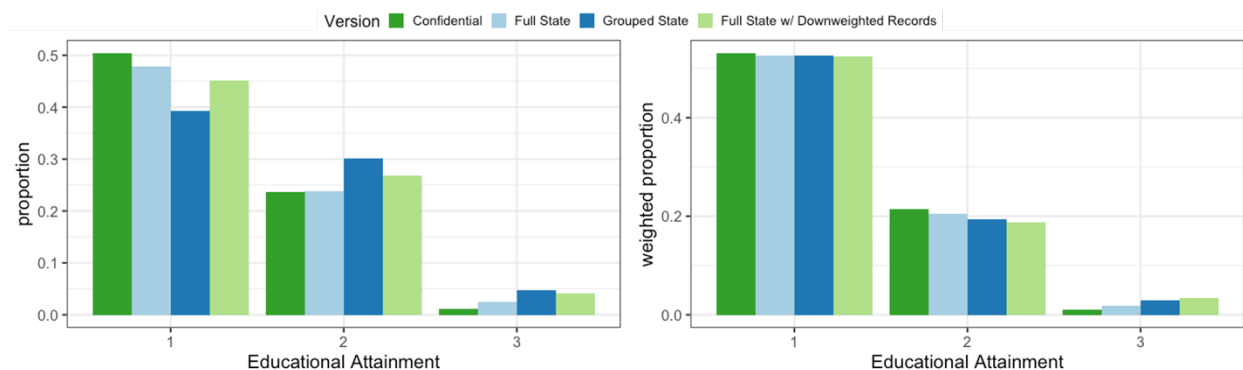


Figure 11. Distribution of Education Categories. Showing synthesis models using all states, grouped states, and down-weighted records. Left panel shows unweighted proportions and right panel shows weighted proportions.

Figure 11 shows the distribution of educational categories for individuals in the same small state previously shown, with three different synthesis methods. The grouped and down-weighting approaches produce counts that differ more from the confidential counts, but the weighted proportions using the individual weights in the data produce similar distributions regardless of synthesis method. This suggests that some of the decrease in utility can be recaptured after the fact by recalibrating the weights to meet desired population totals.

Figure 12 shows the analogous results for the distribution of *Working status*. The weights were not calibrated to this variable, and we see that the weights do not help to make the synthetic

distribution any closer to the confidential distribution. Future work could consider re-weighting to a broader set of targets in order to recapture more utility in the data.

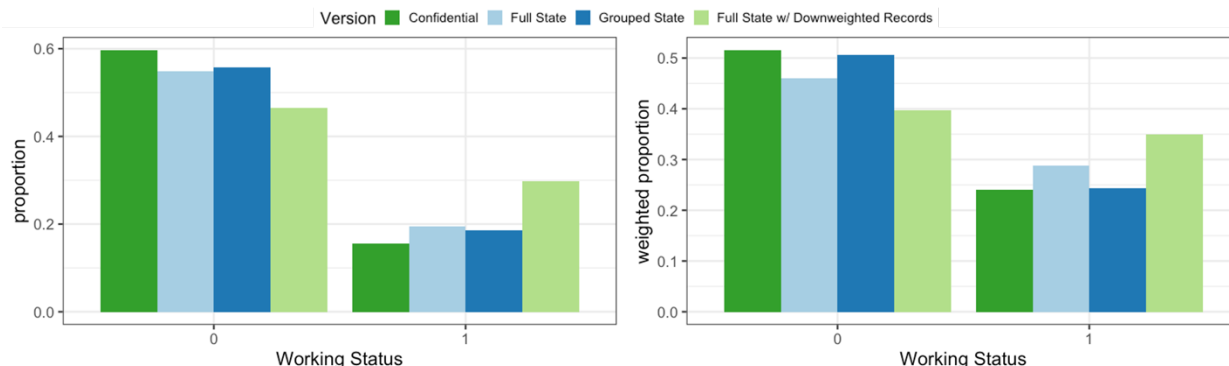


Figure 12. Distribution of Working Status. Showing synthesis models using all states, grouped states, and down-weighted records. Left panel shows unweighted counts and right panel shows weighted proportions.

Measuring Risk

Potential metrics to evaluate disclosure risks can be categorized into identification and attribute disclosure risks. Identity disclosure risk, whether an attacker can correctly identify individual observations in the confidential data using the synthetic data, is a larger concern for partially synthetic data, where only a subset of variables is synthesized (Drechsler, 2011; Hu, 2018). In that case, an attacker can link un-synthesized values with external data to identify records, in the same way as they would attempt to attack other types of published data which contain confidential values. Methods of evaluating identity disclosure risk in partially synthetic data include calculating and reporting the expected match risk, the true match rate, and the false match rate by assuming intruder’s knowledge and behavior (Reiter and Mitra, 2009; Hornby and Hu, 2021) These methods assume certain levels of intruder’s knowledge of databases that contain information about targeted records.

Identification risk still exists for fully synthetic data, though it is measured differently. This is measured through the membership inference attack, which uses a machine learning model to predict whether individual records were part of the confidential data used to generate the synthetic data (Zhang et al., 2022). This concept originated with attempts to identify records used to train machine learning models (Shokri et al., 2017) and has been extended to synthetic data. Very simply, the idea is that if the synthetic data observations are too similar to the confidential observations, an attacker can infer that specific individuals’ data were likely in the confidential data. After identifying records that were likely used to generate the synthetic data, an attacker might go on to try to use the synthetic records to disclose information about additional attributes. This process is termed attribute disclosure.

Attribute disclosure risk, if an attacker can make precise inferences about individual observations in the confidential data using the synthetic data, is a concern for both partially and fully synthetic data. Attribute disclosure typically occurs after identification, though it can also occur through group identification. For example, all synthetic observations matching a candidate individual may share the same sensitive attribute revealing its value without precise identification. The

current data include some sensitive characteristics, such as income, and we will evaluate the inclusion of future sensitive variables.

To compute attribute disclosure risk, we predict sensitive variables in the confidential data using the synthetic data as training data. One can use all non-sensitive variables as predictors and then use distance-based matching, probabilistic matching, and predictive models to generate predictions for the most sensitive variables (Hu, 2019). We will then calculate measures like root mean square error and accuracy to summarize disclosure risk. Both attribute disclosure evaluations and inference membership typically use holdout datasets to improve the interpretation of the risk metrics.

We do not provide results from risk evaluations in this report, but the results from such an evaluation will inform future decisions concerning releasing synthetic data. Given the nature of the LASI, we expect most records will be well-protected because they are a relatively small sample compared with the size of the population. On the other hand, we expect the most significant risk to exist for individuals in households sampled from geographies with very small populations, such as the islands of Lakshadweep.

Finally, because we aim to release household records, each observation in the data will contain significantly more information than a record about a single individual. The combination of the couple and household structure, along with the individual information about each person, makes it so that households quickly become uniquely identifiable in the survey. The task becomes to estimate how identifiable these records will be in the population. Future work will provide concrete estimates of this that will inform release decisions.

5. Conclusion and Next Steps

In this report, we present a first prototype synthetic data set using data from the LASI wave 1. We provide a means of synthesizing the data that preserves the individual, couple, and household structures. We show that the general statistical properties of the confidential data can be reproduced using fully synthetic data. The synthetic data capture aspects such as the univariate, bivariate, and three-way relationships in the data, and they also capture relationships between couples and different individuals in the households. There are areas for improvement, but our approach preserves much of the distribution for the variables included in the prototype.

We propose a novel method for maintaining the couple and household structures, and future work could more formally compare our method to existing methods that attempt to capture structure. We also plan on evaluating different choices for maximum household size to determine the optimal approach that balances capturing household characteristics against the disclosure risk of accurately capturing rare households with many respondents. We will pursue the down-weighting approach further which showed promise to maintain utility while shrinking the contributions from risky records. We have also connected these ideas to measures of disclosure risk and plan to formalize this in future work.

We plan to expand the number of variables, either by including more questions from the LASI or by linking the variables here to additional data to synthesize and release jointly. As we do this, we will need to monitor different parts of the joint distribution of variables to ensure that it

continues to capture the distribution observed in the confidential data. We may also target improvements on variables in the current file which have worse utility, such as ADL or working status.

Moving towards releasing synthetic data will require more considerations of the precise use-cases of the synthetic data, which will inform what variables are included in the data, the target accuracy requirements, and the level of risk mitigation required prior to releasing the synthetic data. The desired uses will also determine whether the data are released publicly or with additional access restrictions. Broadly speaking these decisions require policy decisions and not only technical work. To aid this, we will seek feedback from the user community on the purpose of these synthetic data and the minimum accuracy and privacy protections that are needed before we can release a product.

References

- Benedetto, G., & Totty, E. (2020). Synthesizing Familial Linkages for Privacy in Microdata. Census Bureau Working Paper Number CED-WP-2020-004. <https://www.census.gov/library/working-papers/2020/adrm/CED-WP-2020-004.html>
- Bloom, D. E., Sekher, T. V., and Lee, J. (2021). Longitudinal Aging Study in India (LASI): new data resources for addressing aging in India. *Nature Aging*, 1, 1070-1072. <https://doi.org/10.1038/s43587-021-00155-y>
- Bugliari, D., et al. (2023). RAND HRS longitudinal file 2020 (v1) documentation. Santa Monica, CA: RAND Center for the Study of Aging. <https://doi.org/10.7249/TLA2097-1-v2>
- Chien, S., Young, C., Phillips, D., Wilkens, J., Meijer, E., Angrisani, M., & Lee, J. (2021). Harmonized LASI documentation, version A.2 (2017-2019). Los Angeles, CA: University of Southern California, Center for Economic and Social Research. <https://g2aging.org/>
- Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation* (Vol. 201). Springer Science & Business Media.
- Drechsler, J. and Reiter, J.P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, 55(12), 3232–3243.
- Hornby, R. and Hu, J. (2021). “Identification Risks Evaluation of Partially Synthetic Data with the IdentificationRiskCalculation R package, *Transactions on Data Privacy*, 14:1, 37-52.
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P. P. (2012). *Statistical disclosure control* (Vol. 2). New York: Wiley. <https://doi.org/10.1002/9781118348239>
- Hu, J. (2018). Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data. *Trans. Data Priv.*, 12, 61-89.
- Hu, J., Reiter, J. P., & Wang, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data.
- Juster, F. T., & Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, 30, S7–S56.
- Lee, J., Phillips, D., Wilkens, J., and Gateway to Global Aging Data Team (2021). Gateway to Global Aging Data: Resources for Cross-National Comparisons of Family, Social Environment, and Healthy Aging. *Journals of Gerontology: Social Sciences*, 76:S1, S5-S16. <https://doi.org/10.1093/geronb/gbab050>
- Little, R. J. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, 9(2), 407.

NIA (2007). Growing older in America: the Health and Retirement Study (NIH Publication No. 07-5757). Bethesda, MD: National Institute on Aging.

Nowok, B., Raab, G.M. and Dibben, C., 2016. synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software*, 74, pp.1-26.

Nowok, B., Raab, G.M., Snoke, J. and Dibben, C., 2016. Synthpop: generating synthetic versions of sensitive microdata for statistical disclosure control. *R package version*, pp.1-3.

Perianayagam, A., et al. (2022). Cohort Profile: The Longitudinal Ageing Study in India (LASI). *International Journal of Epidemiology*, 51(4):e167-e176. <https://doi.org/10.1093/ije/dyab266>

Raab, G. M., Nowok, B., & Dibben, C. (2016). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3), 67-97.

Raab, Gillian M., Beata Nowok, and Chris Dibben. "Assessing, visualizing and improving the utility of synthetic data." *arXiv preprint arXiv:2109.12717* (2021).

Raghuathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1), 1.

Reiter, J.P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21(3), 441–462.

Reiter, J. P., & Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1(1).

Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, 53(4), 1475-1482. <https://doi.org/10.1016/j.csda.2008.10.006>

Rocher, L., Hendrickx, J.M. & de Montjoye, YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* **10**, 3069 (2019). <https://doi.org/10.1038/s41467-019-10933-3>

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

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.

Zhang, Z., Yan, C., & Malin, B. A. (2022). Membership inference attacks against synthetic health data. *Journal of biomedical informatics*, 125, 103977.