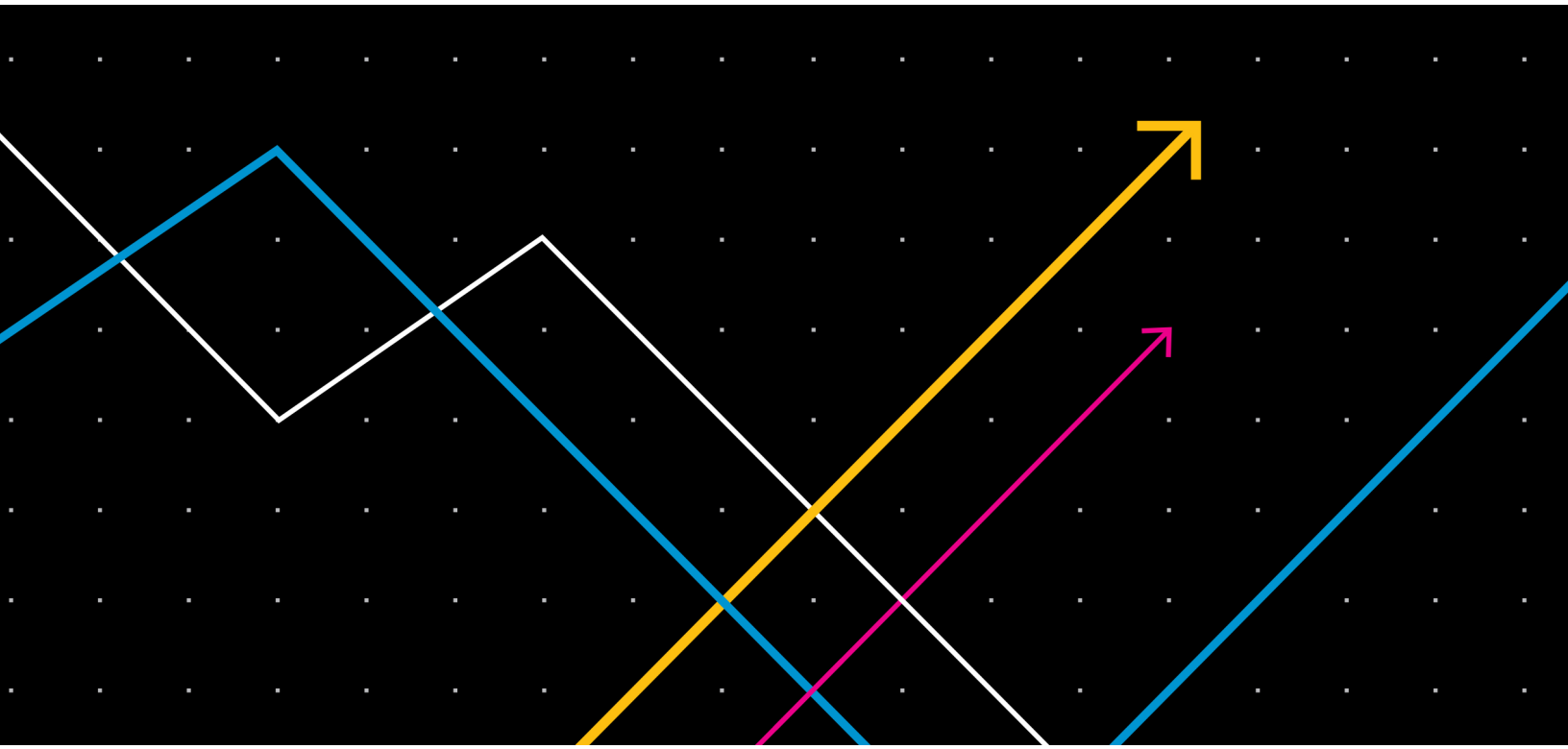


RESEARCH REPORT

Generating a Fully Synthetic Human Services Dataset

A Technical Report on Synthesis and Evaluation Methodologies

Madeline Pickens
URBAN INSTITUTE

Jennifer Andre
URBAN INSTITUTE

Gabriel Morrison
URBAN INSTITUTE

August 2023

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

Contents

Acknowledgments

This report is a product of the Urban Institute's Racial Equity Analytics Lab, which operates with the generous support of the Ballmer Group, the Bill & Melinda Gates Foundation, the Salesforce Foundation, and Urban's general support donors. Lead funding for this report was provided by the Bill & Melinda Gates Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

The authors thank our other project team members, Dr. Claire McKay Bowen and Alena Stern, as well as the following external partners, without whom this work would not have been possible:

- Geoffrey Arnold (Allegheny County Office of the County Manager)
- Kathryn Collins (Allegheny County Department of Human Services)
- Joanne Foerster (Allegheny County Office of the County Manager)
- Robert Gradeck (Western Pennsylvania Regional Data Center)
- Ross Reilly (Western Pennsylvania Regional Data Center)
- David Walker (Western Pennsylvania Regional Data Center)

The authors also thank Aaron R. Williams, who generously provided invaluable feedback that greatly improved this work.

Generating a Fully Synthetic Human Services Dataset

The Department of Human Services (DHS) in Allegheny County, Pennsylvania, serves one in five residents of the county every year through child welfare services, behavioral health services, aging services, developmental support services, homeless and housing supports, and family strengthening and youth supports. In the process, data are collected about these services and the population using them. Because of the dataset's sensitive nature, DHS has not widely shared the data at an individual level. An alternative solution is using synthetic data generation to create pseudo or fake data to replace the real dataset—allowing the data to be publicly shared and helping stakeholders, including researchers, service providers, and members of the public, understand these populations better.

A team at the Urban Institute collaborated with the Allegheny County DHS and Western Pennsylvania Regional Data Center (WPRDC; hereafter "county partners") to create a fully synthetic version of the 2021 Integrated Services dataset (hereafter "2021 Synthetic Integrated Services"), meaning we entirely replaced the underlying records that track the use of these services with statistically representative pseudo-records. This synthetic disaggregated (individual level) dataset preserves many of the features and statistical properties of the confidential human services data while also mitigating some of the privacy threats to individual participants reflected in the data. This report details our data synthesis and evaluation methodologies.

Data Privacy Background and Definitions

Modern advancements in technology have enabled the massive growth in the volume and availability of data, including sensitive data about individuals. Data privacy refers to the right of individuals to control the disclosure of sensitive information about themselves (Fellegi 1972). Synthesizing data is just one technique in a wide range of statistical methods proposed and implemented by data privacy researchers to balance the tradeoffs between data quality and data privacy, broadly referred to as statistical disclosure control methods (Bowen 2021). This section provides a brief overview of key terms related to disclosure risks and synthetic data.

Disclosure Risks

Even if a dataset is deidentified, or stripped of variables that explicitly identify entities in the data, the release of disaggregated, individual-level data creates disclosure risks for those entities. More specifically, malicious actors can use supplemental datasets to attempt to reidentify individuals and disclose sensitive information about them. The growing availability of public administrative datasets, the ease of access to high-powered computing resources, and the advancement of statistical methods exacerbate these risks.

There are three main types of disclosure risk: identity disclosure, attribute disclosure, and inferential disclosure. Identity disclosure refers to the association of a specific individual with a released data record. For example, a data user could determine the identity of a specific individual in a dataset by cross-referencing their date of birth with a birthday announcement in a local newspaper. Attribute disclosure refers to the determination of an individual's characteristics based on other information in the released data. For example, a data user can determine that a 50-year-old individual is covered by Medicaid if other information in the data indicates that all individuals above age 50 in the dataset are covered by Medicaid (Bowen, Williams, and Pickens 2022). Inferential disclosure occurs if the data intruder predicts the value of some characteristic from an individual more accurately with the public data or statistic than would otherwise have been possible. For example, if a public homeownership dataset reports a high correlation between the purchase price of a home and total income, a data adversary could infer someone's income based on purchase price listed on Redfin or Zillow (Bowen, Williams, and Pickens 2022).

Fully Synthetic Data

Synthetic data are pseudo-records generated from statistical or predictive models that can be representative of the confidential data and used for valid analysis. To generate partially synthetic data, only some variables are synthesized, so there remains a one-to-one mapping between the synthetic and underlying confidential records. In contrast, to generate fully synthetic data, all variables are synthesized and there is no one-to-one mapping between synthetic and confidential records. Even so, the synthetic data generation process could still perfectly replicate the original records.

Developing a Statistical Data Privacy Method

Generally, there are three phases of developing a statistical data privacy method:

1. *Preprocessing* the data and creating a gold standard dataset (GSDS), the ideal, cleaned version of the original, confidential data, to serve as an input in the privacy step.
2. The *privacy step*, which refers to actually applying methods to enhance privacy. In this case, the modeling and synthesis of the GSDS comprise our privacy step.
3. *Postprocessing* the output of the privacy step to ensure consistency and quality of the dataset. Adjustments in this step can add additional privacy protections.

These steps are often iterative and informed by the privacy and utility thresholds defined at the beginning of the process.

Allegheny County Human Services Data

Administrative data, or data collected about the operations of an organization, can help stakeholders understand the complexities of organizational performance. The Allegheny County DHS collects administrative data about service usage for the purpose of care coordination, case management, and quality improvement efforts. This section provides an overview of current human services data access options, the limitations of these options, imagined use cases, and key priorities for the 2021 Synthetic Integrated Services dataset.

Current Data Access Options and Limitations

The Allegheny County DHS currently releases a version of its human services data on the WPRDC open data website (Allegheny County DHS 2022). While these data are released publicly, they are aggregated, or grouped on specific features, to protect individual privacy. The level of aggregation is at the municipality scale, meaning the DHS releases the total number of people in a municipality receiving a service. Moreover, DHS suppresses (i.e., does not report publicly) values when fewer than six individuals from a municipality receive those services. Although aggregating and suppressing the data can protect individual privacy, these measures limit potential applications of the service use data. Interactions between usage of different services, as well as the repeated nature of the service distribution, are not captured. For example, the existing public data would not allow a user of the data to understand whether a municipality's total service count reflects many residents receiving a single service, or a small minority of residents receiving many distinct services.

Some types of analyses are impossible with the aggregated and suppressed data. For example, consider a researcher studying the effect of service provision on upward mobility. This researcher could investigate whether an individual receiving income support and other services at the beginning of the year was more or less likely to continue to receive income support near the end of the year relative to another individual who only received income support. Such an analysis would be impossible with the publicly available data, which only includes annual service usage and does not include data on cross-service usage. Similarly, the public municipality-level aggregates would be insufficient for a researcher hoping to analyze use of a service by demographic features—for example, to determine whether specific age cohorts could be targeted for an intervention.

The publicly released aggregated data contrast with the Allegheny County DHS's private disaggregated data, which is composed of individual-level records with information about a particular service recipient and the services they received. While the public aggregates may be sufficient for some use cases, many research projects may require individual-level data to explore interactions between usage of different services or the repeated nature of service distributions over time. Researchers may request access to the confidential disaggregated data through a formal request process with the Allegheny County DHS. However, it may be difficult to make such requests without knowing the disaggregated data structure and ranges of values, which we discuss further in the following section.

Synthetic Data Use Cases

Releasing disaggregated synthetic data can simplify the research process. The synthetic data would have a similar structure (i.e., the same variables and range of values) to the confidential data, so researchers can analyze it without the administrative burden of the formal data access petition process. Data users also do not need to worry about working with the data in a secure computing environment or physically traveling to a secure enclave. Additionally, since the synthetic data captures many relationships in the confidential data, researchers can explore general relationships between types of services received, trends over time, and associations between age and service receipt.

Even if data users find that their desired analysis is still not possible with the synthetic data, its release can still improve the research process. Historically, the DHS has found that researchers would request access to the entire dataset to understand its structure and ensure they had all necessary data, whether or not they needed all of it to conduct their research. By releasing the synthetic dataset with a structure that is very similar to the confidential dataset, researchers can be more specific in their requests and the DHS has less reason to permit access to the entire confidential dataset. While waiting

for access to the confidential dataset, researchers can also rely on the synthetic dataset as training data. They can run and debug their analyses on the synthetic data and have functional code that is ready for use once access to the confidential data is granted.

Synthetic Data Priorities

To ensure privacy protections remain robust, not all elements of the confidential dataset can be preserved through the synthesis process. Therefore, developers of synthetic data must prioritize certain features and properties of the confidential data that are most important to preserve in the synthesis, and these features should align closely with typical use cases of the data. The most effective synthetic dataset mimics these high-priority features.

We discussed these use cases and priority features with the county partners early in this project and returned to these discussions regularly. We learned that end users of the record-level confidential data are typically researchers interested in the integration of service usage over time and the interactions between service usages. In addition to the importance of preserving these relationships, the County emphasized the importance of preserving key demographic features of service recipients including age, race, ethnicity, and gender. The County considered other features of the data, such as service recipient living situation, marital status, and education, to be less important to preserve when evaluating the synthetic data.

Data Synthesis Methodology

To construct our synthetic dataset, we preprocessed the data into a gold standard dataset (GSDS), the ideal, cleaned version of the original, confidential data, that served as an input for our privacy step. We modeled the GSDS in two stages to ensure we were capturing all the important aspects of the data in our privacy step, which created a fully synthetic dataset. Finally, we postprocessed the data to make sure that the synthetic product was consistent and matched real-world constraints.

Data Structure and Features (Preprocessing Step)

To synthesize the data, we needed to first consider the structure most well-suited to the use cases described to us by our county partners. We then restructured the data into a single GSDS and included the exact variables we planned to synthesize.

Dataset Background

Allegheny County DHS provided the Urban team with data tracking individual's usage of 22 services. Note that we refer broadly to "services" throughout this report to refer to the types of interactions with DHS included in the data. However, some types of interactions like "overdoses" or "individuals in the Allegheny County jail" are better described as "incidents" of these types of events and are denoted with an asterisk. The full list of services is as follows:

- Individuals receiving DHS services
- Families receiving child welfare services
    - Children receiving child welfare services
    - Parents receiving child welfare services
    - Children in care
- Children receiving DHS funded out-of-school programs
- Children attending early childhood programs managed by DHS
    - Children receiving early intervention services
- Individuals receiving family strengthening programs

- Individuals receiving homelessness and housing services
    - Individuals identified as homeless
- Individuals receiving intellectual disability services
- Individuals receiving mental health services
    - Individuals receiving services for a mental health crisis
- Individuals with an involuntary commitment
- Individuals receiving substance use disorder services
- Individuals receiving income supports
- Individuals in the Allegheny County jail*
- Older adults receiving services
- Homicides*
- Overdoses*
- Suicides*

Data for each service are stored separately within the county's integrated data warehouse but are linked by a common ID for each individual who receives a service. For each service, we received a dataset with the following variables:

- Client identifier
- Year of event
- Date of event
- Geographic area
- Date of birth
- Date of death
- Gender
- Gender identity
- Sexual orientation
- Legal sex

- Race
- Ethnicity
- Living arrangement
- Employment status
- Marital status
- Education level
- Veteran flag

The county provides many of the services on an ongoing basis, but it tracks service usage on a monthly level as a binary variable indicating whether a recipient did or did not receive a service at least once in a given month. The "date of event" variable took the value of the last day of each month. "Year of event" exclusively took the value 2021 because all data were from 2021. Demographic variables for each individual remain consistent across services and are current to when the data were pulled in June 2022. Also note that many records have missing values for some variables.

Creating the Gold Standard Dataset

The first step in any data synthesis is to create and define a GSDS that can serve as the raw data for synthesis and the benchmark data for evaluating quality, as the data collection process can be imperfect and introduce messiness and inconsistencies to the raw data. As Wooley and colleagues (2020) note, "Synthetic data are only as good as the raw datasets from which they are imputed," so it was crucial to ensure that the gold standard dataset that we defined preserved key features of the data.

Our first step in creating this dataset was to merge the separate files for each service and create a master file that contained every individual and their interactions with the county system throughout 2021. We also converted the "date of birth" variable to age (calculated as of the end of 2021) and "date of event" to the "month" in 2021 when the service was received. Finally, we created an explicit "service" variable to reflect the service provided.

Structuring the Gold Standard Dataset

We considered two major questions when structuring the merged data: which structure would best enable fitting models and imputing the synthesized data, and which structure would reflect the most

useful output as a dataset product for users. We revisit the first question when discussing our two-stage synthesis procedure in Synthesis Process section. Regarding the second question, we returned to the use case defined by the county, which emphasized tracking interactions between service usage and understanding service usage over time. We decided that our GSDS structure would be on the service-month level, with each row reflecting one service received by one individual in a given month. A high-level example of this structure is shown below in table 1 for five hypothetical individuals receiving a mix of services. In this example, the individual with ID 1 receives service 1 in January, February, and November, and the individual with ID 4 receives service 1 in June and service 5 in July. Demographics remain consistent across each ID.

TABLE 1

Example of the Gold Standard Dataset Structure

Mock data for illustrative purposes

ID	Month	Age	Gender	...	Service
1	January	42	M	...	1
1	February	42	M	...	1
1	November	42	M	...	1
2	March	6	F	...	2
2	April	6	F	...	3
3	February	17	F	...	4
4	June	28	M	...	1
4	July	28	M	...	5
5	December	65	F	...	1

Source: Mock data illustrating confidential dataset structure

Notes: We list the services as integers for simplicity, but in the actual gold standard dataset and synthetic dataset, the service data is a string with that encodes one of the 22 services.

Variable Exclusions

Some of the demographic variables in the confidential data were very sparse, which posed a problem for both quality and privacy of the synthetic output. Specifically, the sparsity would make the variables difficult to model well, and the fact that so few individuals were associated with values for those variables meant that re-identifying the individuals would be potentially much easier. For these reasons, we chose to exclude the following variables from the synthetic output:

- Gender identity
- Sexual orientation

- Legal sex
- Employment status
- Veteran status
- Date of death

We also excluded the "geographic area" variable from the synthetic output. While this variable was more complete than the other excluded variables, it posed separate problems for the quality of our synthesis because of its balance issues. Although the data contained 129 geographic areas, Pittsburgh made up 21 percent of the records and the next most-represented area made up only 3 percent of the records. This imbalance meant that, in a similar way to the sparse variables, it would be difficult to accurately synthesize this variable. Additionally, successfully capturing this variable might require sacrificing other elements of the data quality that were more relevant to the prioritized use case (notably, service usage over time). A successful synthesis of this variable would also pose privacy issues because it would be potentially easier to re-identify individuals associated with smaller locations and disclose sensitive information about them. Smaller geographies might even need to be suppressed entirely to align with county policy. With these issues in mind, the costs did not outweigh the potential benefits of synthesizing "geographic area."

Although we did not synthesize the seven variables described above, we did include them as variables populated with missing values in our synthetic dataset. We added these variables to ensure that the structure of the synthetic data is consistent with the confidential data and informative for users who may want to request access to the confidential data.

Synthesis Process (Privacy Step)

The structure of the GSDS format was closely aligned with the target use case, but it was not ideal for modeling service usage over time. Because of the sequential nature of our synthesis, where only previously synthesized variables can be used as predictors, if we trained our models on the GSDS format, we would not capture the relationship between receiving multiple services simultaneously or single or multiple services over time. In an attempt to better capture these relationships, we synthesized the data in two stages in order to arrive at our final GSDS format.

Stage One: Synthesizing Service Usage Counts

In the first stage of the synthesis, we converted the GSDS dataset to include only one row per individual. The dataset included additional columns for each service, reflecting the total number of months in 2021 that individual received the service. Table 2 shows how the data from Table 1 could be aggregated to reflect that the individual with ID 1 received service 1 three times in 2021, while the individual with ID 4 received service 1 only once.

TABLE 2

Example of Service Counts Dataset Structure

Mock data for illustrative purposes

ID	Age	Gender	...	Service 1	Service 2	Service 3	...
1	42	M	...	3	0	0	...
2	6	F	...	0	2	1	...
3	17	F	...	0	0	1	...
4	28	M	...	0	1	0	...
5	65	F	...	1	0	0	...

Source: Mock data illustrating confidential dataset structure.

Our goal in this stage of the synthesis was to create a synthetic file that closely matched distributions of service usage by demographic (particularly age, gender, race, and ethnicity) and the counts of service-months (total months in 2021 in which each service was received).

To synthesize this data, we used *tidysynthesis*, an R package that generates synthetic data using the underlying design philosophy, grammar, and data structures of the *tidyverse* and *tidymodels* programming packages by Posit, formerly RStudio (Williams 2022). This package will eventually be made publicly available following additional user testing and review.

We used decision tree and regression tree models to impute the synthetic data. These models use predictor variables to sort observations of an outcome variable into groups of observations with relatively similar characteristics that have predictive power. We chose to focus on tree-based models for this data because, in addition to being computationally simple and flexible, these models have been shown to perform well for categorical data (Drechsler and Hu 2021), which comprises most of this dataset. Decision and regression tree users can specify model hyperparameters that control how deep these trees will grow, or how many times the data will be split before predictions are made based on the resulting groups (Therneau and Atkinson 2022). We estimate models for each variable sequentially. This method uses previously synthesized outcome variables as predictors, so the order in which

variables are synthesized has a large impact on the synthesis results because variables synthesized earlier in the sequence are subjected to less propagated modeling error. Rather than using the traditional prediction functions in decision and regression trees like mean or mode, we sample uniformly from the final node, which provides more variation in the synthetic output.

We tested a variety of synthesis specifications by varying modeling hyperparameters, variable order, and mid-synthesis constraints (logical conditions imposed on variable values), resulting in 77 versions of the synthetic dataset.

Our selected synthesis started with a random sample of the "income supports" variable (the most common service) and then synthesized the variables in the following order: race, ethnicity, gender, age, every other service in order of most- to least-received, living arrangement, marital status, and education level.

During this phase of the synthesis, we applied constraints to the age variable to ensure that the services assigned to an individual were realistic. More specifically, our county partners provided us with eligibility information for each of the services, and we used the information to ensure that synthetic records matched these constraints. As an example, Allegheny County provided many services targeted toward children only to people 18 and under, so we ensured the maximum age of a record receiving those services was also 18. We also checked that service counts were realistic in the domain context, not exceeding the maximum number of services received by one person in the confidential data. Some "services" (homicides, suicides, overdoses, and mental health crises) only appear in the confidential data once per individual, so we also ensured that this constraint was reflected in the service counts.

Stage Two: Generating Monthly Service Distributions

In the second stage, we took the synthesized individual service counts and applied an algorithm to convert these counts to a distribution of assigned months for each service. Previously, each service was represented as one column of data populated with service receipt counts for a synthetic individual.[1] The output of the algorithm would be, for each service, 12 columns populated with binary values indicating whether or not that service was received in the given month. Because this results in 264 columns (12 months multiplied by 22 services) for each synthetic individual in the data, it was not computationally feasible to model and impute the confidential version of this data as we did in stage one. Instead, for each service recipient in the data, and for each month in the year, the algorithm performed the following steps:

1. For a given service, do the following:
    a. Check if a synthetic recipient "must" receive that service. For example, if the algorithm is working on the January column for that service and a synthetic record receives the service in all 12 months, that record must receive that service in January.
    b. Check if a synthetic recipient cannot receive a service. For example, if a synthetic service recipient has received the service in January, and it only receives the service once over the full year, it cannot receive that service in any subsequent months.
    c. If the service receipt is not determined by steps 1 or 2, determine whether a service is received using a probability stored in a look-up table.
2. Repeat for all other services.

To calculate the probabilities from step c, we created a separate probability for each unique pairing of service (s), count of times that service was received (t), and month the service was received (m). We calculated the probability that a service would be received in each month as follows:

$$p_{s,m,t} = \frac{\text{Individuals receiving service } s \text{ a total of } t \text{ times during month } m}{\text{All individuals who received service } s \text{ a total of } t \text{ times}}$$

In other words, the probability of whether or not a service was received is calculated as the total number of individuals in the confidential data who had received a given service t months and *during the month in question*, divided by the total number of people who received the service t times. For this calculation, we ensured that we only considered records in the confidential dataset that would not have met checks in steps 1 and 2 of the algorithms above for the month in question.

Once we had assigned months to the service recipients, we reevaluated quality and privacy, particularly regarding the distribution of services by month and the demographic breakdowns of service counts. We selected a final candidate synthesis that performed well across all these dimensions. The specific privacy and quality elements that we evaluated are discussed in the Evaluating and Selecting a Synthetic Dataset section.

Postprocessing

Once we had assigned months to the service recipients using the Stage Two process described above, we reshaped the data to match table 1, the original gold standard format.

We also performed additional post-processing to account for unrealistic values that appeared in synthesis output that could not be addressed by mid-synthesis constraints. We converted unrealistic values for marital status and education level for certain individuals under age 18 to "unknown." For example, we applied this postprocessing to children under 16 with an education level higher than high school or children under 18 with a marital status other than "single never married" or "unknown." We found that this method of dealing with unrealistic values provided the best balance between ease of application and retaining the other features of the synthetic data relative to other options.

Evaluating and Selecting a Synthetic Dataset

As mentioned previously, we ultimately generated 77 synthetic datasets. The selected 2021 Synthetic Integrated Services dataset is the version that best balanced data quality and data privacy for this use case. We worked closely with the county partners to determine their priorities for data quality and to ensure that our assessments of data privacy aligned with county privacy policies.

Evaluating Synthetic Data Quality

To evaluate data quality, often referred to as data utility, we sought to determine how well the synthetic dataset captured and mimicked the underlying properties of the confidential data. It is not possible (or desirable, from a privacy perspective) to perfectly capture all features of the confidential data in the synthetic data. We therefore selected key properties to prioritize when making these utility assessments. We worked closely with the county partners to identify salient utility metrics for this context and ultimately compare these across candidate syntheses to select a final synthetic dataset.

Utility Considerations

Early in the project, we discussed common data use cases with the county partners to determine high priority features and properties to preserve in the synthetic data. As previously noted, our discussion with county partners led us to prioritize these aspects of the confidential dataset ranked from most to least important:

1. Service usage over time and the interactions of between different services
2. Key demographic features: age, race, ethnicity, and gender
3. Other demographic features: living situation, marital status, education level.

Utility Metrics

We used a range of general utility metrics to evaluate the quality of the synthetic data. General utility metrics measure the distributional similarity between the confidential and synthetic data without

assuming a specific type of analysis use case. We discussed findings with our county partners to ultimately select a synthetic dataset that had high performance across these metrics relative to other implicates.

Categorical relative frequencies. We compared relative frequencies of each value or combination of values of categorical variables between the synthetic and confidential data to measure how successfully the data synthesis captured the distributions of these values. For greatest synthetic data utility, we wanted to see very similar proportions of each level of each variable in both datasets (e.g., the proportion of records with race "Black/African American" is approximately the same in both the confidential and synthetic data; Bowen et al. 2020).

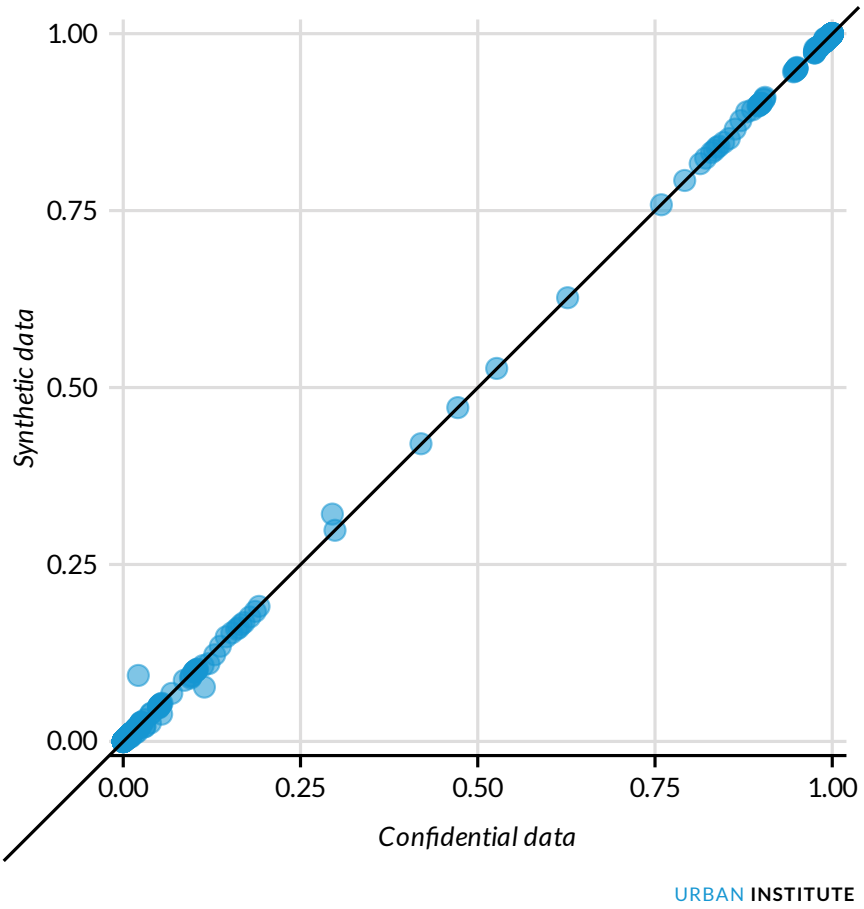
We explored one-way relative frequencies (i.e., each variable independently), three-way joint relative frequencies (i.e., combining gender, race, and ethnicity), all demographics joint relative frequencies (i.e., combining gender, race, ethnicity, marital status, education level, and living arrangement), and all-way joint relative frequencies (i.e., all variables combined). In each case, we evaluated the minimum, maximum, mean, and median differences, absolute value differences, and proportion differences among each value or combination of values between the synthetic and confidential data.

The synthetic data candidates with greatest utility had smaller differences in these relative frequencies. In general, these differences were smallest for one-way comparisons and became larger as variable combinations increase to all-way relative frequencies, reflecting the difficulty of maintaining high-quality relationships across multiple variables.

We also plotted these relative frequencies for each value or combination of values, with the confidential proportion on the x-axis and synthetic proportion on the y-axis. Plots with points adhering closely to the 45-degree diagonal line indicate greatest synthetic data quality. For example, figure 1 demonstrates that our selected confidential dataset had high data quality for one-way relative frequencies using the selected synthetic dataset. The close adherence of points to the 45-degree line demonstrates that our synthesis effectively captured the one-way frequencies from the confidential dataset.

FIGURE 1

Data Synthesis Effectively Captured One-Way Relative Frequencies from the Confidential Dataset

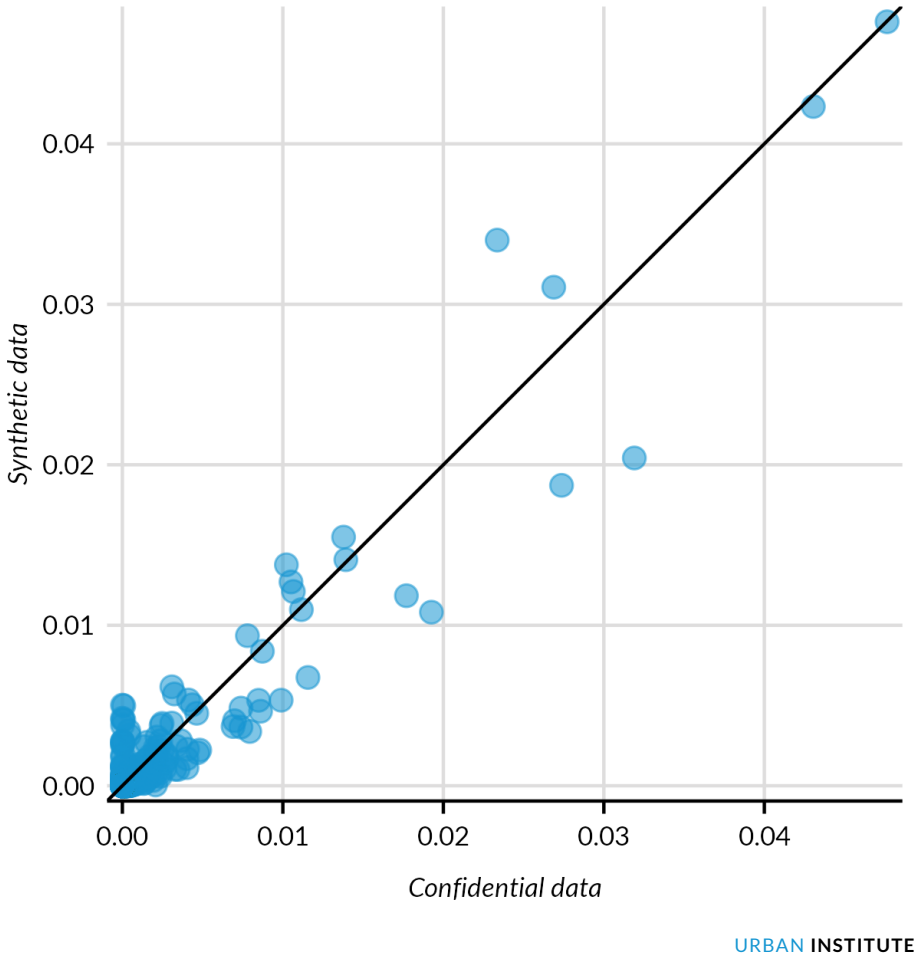


Source: 2021 DHS confidential data and 2021 Integrated Services data.

Figure 1 shows that the synthetic dataset captures one-way relative frequencies very well. However, we are also interested in exploring how well the synthetic dataset maintains relationships between variables. We can use comparisons of multiway relative frequencies to explore how well the synthetic dataset captures relationships between multiple variables. Figure 2 demonstrates how well our synthetic dataset captured combinations of all variables from the confidential data, again comparing relative frequencies of these variable combinations between the two datasets. These relationships are more difficult to maintain, and we do see greater dispersion from the 45-degree line in figure 2.

FIGURE 2

All-Way Relative Frequencies Demonstrate Greater Differences across All Variables between the Confidential and Synthetic Datasets



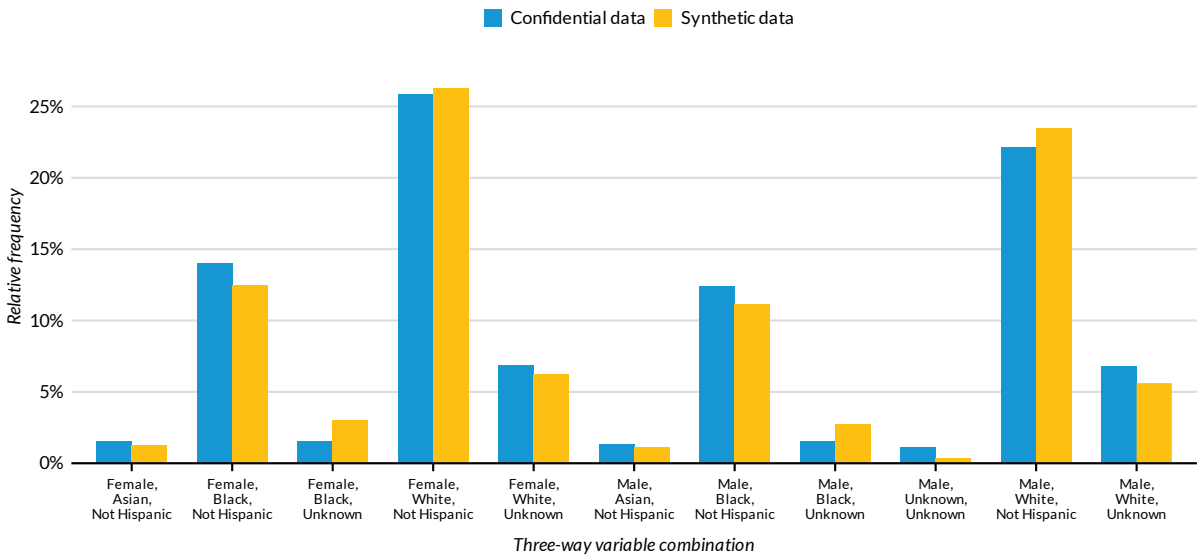
Source: 2021 DHS confidential data and 2021 Integrated Services data.

Rather than look at relative frequencies of single variables or combinations of all variables, we also examined the performance of the synthetic data in capturing subsets of variables of particular interest to us. For instance, figure 3 shows the differences in three-way relative frequencies for gender, race, and ethnicity using the selected synthetic dataset. In this case, we use grouped bar charts to compare the relative frequencies of these combinations between the confidential and synthetic datasets. The figure is filtered to show only combinations of race, ethnicity, and gender comprising at least 1 percent of records in the confidential data. Overall, the relative frequencies from the selected synthetic dataset are very similar to those from the confidential data. However, there are some small differences. For example, male or female residents who are Black and non-Hispanic are slightly underrepresented in the synthetic data, while male or female residents who are white and non-Hispanic are slightly

overrepresented in the synthetic data. Much of the variation in demographics could be because of errors in sampling rather than errors in modeling, as we uniformly sample from the final fitted tree nodes to arrive at our synthetic value.

FIGURE 3

Synthetic Dataset Preserves Relative Frequencies of Key Demographic Combinations



URBAN INSTITUTE

Source: 2021 DHS confidential data and 2021 Integrated Services data

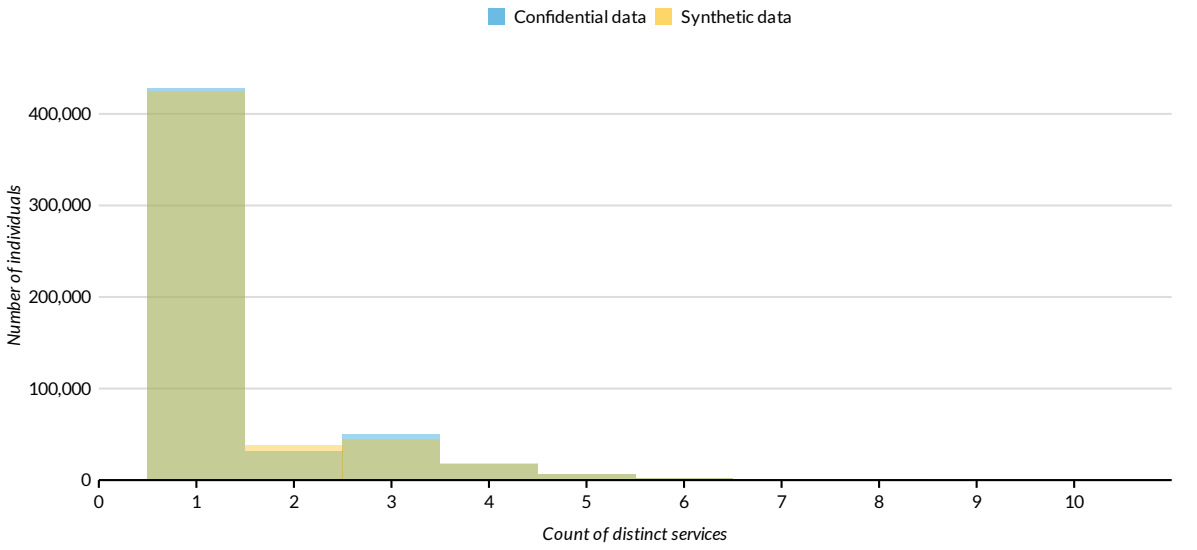
Individual-level service counts. To measure how successfully the data synthesis captured service utilization, we compared the distributions of total service-month counts and distinct service counts for individuals between the confidential and synthetic data. A service-month reflects a specific month in which an individual receives a specific service (e.g., one service for twelve months and a different service for eight months equals 20 total service-months). Distinct service counts reflect the number of distinct services received by individuals for any number of months (e.g., one service for 12 months and one service for 8 months equals two distinct services).

In both cases, significant overlap between these distributions reflects relatively higher data quality. The histogram of service-months counts (not displayed here), shows significant overlap in the synthetic and confidential distributions. The synthetic dataset captures a large spike at 12 service-months, which reflects the most common service receipt pattern—income supports for 12 months, and no other services. There are smaller spikes at other multiples of 12 service-months, reflecting individuals who receive multiple services in each month of the year.

Figure 4 shows the histogram of distinct service counts, again with significant overlap (green) between the confidential (blue) and synthetic (yellow) distributions of individual-level distinct service counts. It is very common for individuals to have received only one distinct service in this period, though some individuals do receive multiple distinct services. Overall, the selected synthetic dataset very effectively captures both individual-level total service-months and individual-level counts of distinct services received.

FIGURE 4

Individual-Level Counts of Distinct Services in the Synthetic Data Closely Mirror Those in the Confidential Data



URBAN INSTITUTE

Source: 2021 DHS confidential data and 2021 Integrated Services data

Monthly service counts. To measure how successfully the data synthesis captured service utilization across months, we compared the counts of total services by month and distinct individuals receiving services by month between the confidential and synthetic data. For total services by month, an individual receiving multiple services is counted several times. For distinct individuals receiving services by month, an individual receiving multiple services is counted only once. We also explored these counts by month separately for each service.

These metrics indicated how well the synthetic dataset captures usage of different services and service utilization trends over time, with small differences indicating a relatively high-quality synthetic data candidate. For the selected synthetic dataset, in the case of the more common services (e.g., income supports), monthly counts tracked the confidential very well overall. While there are some

larger differences in counts for the less common services, the monthly counts from the synthetic dataset still generally capture the shape of the trends across time in the confidential data. For example, the synthetic data shows a drop in use of the early childhood programs in July, corresponding with a similar drop in the confidential data.

Summary of multiservice recipients. In addition to exploring how well the synthetic dataset captured receipt of individual services and services over time, we also explored service combinations. Specifically, we examined how well the synthetic data captured individuals receiving pairs of services throughout the year. We first created a matrix of all possible pairwise service combinations and counted the number of individuals who received both of those services in 2021, for both the synthetic and confidential datasets. We then found the absolute percent difference between these counts and explored these differences. A smaller absolute percent difference indicates that the synthetic data more successfully captures the combination of services.

The magnitude of the absolute percent difference varies based on the frequency of the service combination in the confidential data. The differences tend to be smaller for service combinations that are more common, and larger for service combinations that are less common. For example, 10 percent of individuals in the confidential data receive the combination of both mental health services and income supports, and for this combination of services the absolute percent difference between the counts of recipients in the synthetic and confidential dataset is 2 percent. In contrast, fewer than 0.1 percent of individuals in the confidential data are parents receiving the combination of both the child welfare service and family strengthening programs, and for this combination of services the absolute percent difference between the counts of recipients in the synthetic and confidential data set is 35 percent. Overall, the median absolute percent difference for service pairs that are received by at least 0.5 percent of individuals in the confidential data is 5.2 percent. The median absolute percent difference for service pairs that are received by fewer than 0.5 percent of individuals in the confidential data is 34.4 percent. This implies that our synthesis does fairly well at capturing service pairs for more common services but much worse for rarer combinations of services.

Age summary statistics and distributions overall and by service. To measure how successfully the data synthesis captured the distribution of age, we compared the values of summary statistics like minimum, maximum, mean, median, variance, standard deviation, skewness, and kurtosis between the synthetic and confidential data. Small differences between the datasets would indicate a relatively high-quality synthetic data candidate version (Bowen et al. 2020). We also used density plots to compare the overall distributions of the age variable between the synthetic and confidential datasets. Significant overlap in

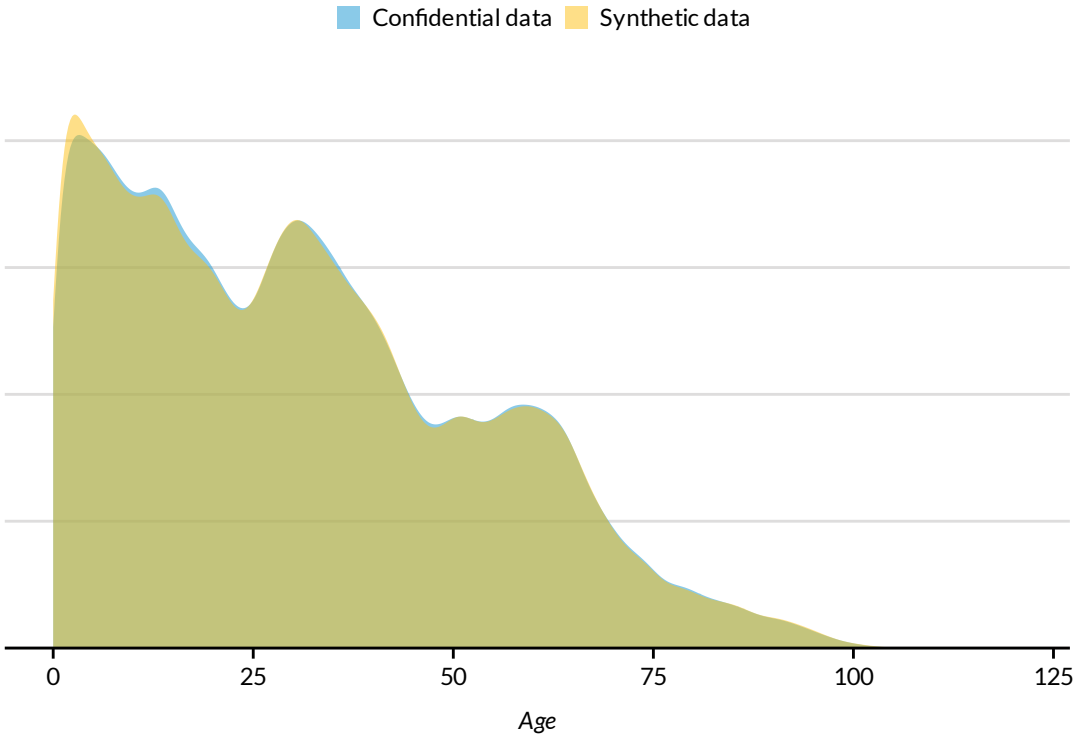
the two distributions indicate a high-quality synthetic data candidate. The statistical moments and distribution of age in the synthetic dataset are very similar to those in the confidential dataset.

Our initial exploratory data analysis revealed meaningful relationships between age and services, especially for services targeted for populations of a specific age range. Therefore, we also explored these summary statistics and age distributions separately for each service and compared these between the synthetic and confidential datasets. As with service receipt counts by month, the selected synthetic dataset shows higher quality for the more common services. While the distributions do not perfectly overlap for the less common services, the synthesis still managed to capture the general shape of the distribution.

For example, figure 5A compares the confidential and synthetic age distributions for recipients of the income supports service using the selected synthetic dataset. The distributions are very closely aligned, with just small deviations for some of the younger ages. This was one of our best-performing services, which is unsurprising both because income supports service was the most commonly received service and was synthesized earlier in the synthesis order, meaning that there was less error propagated through to the synthesized values.

FIGURE 5A

Synthetic Data Aligns Very Closely with the Age Distribution for the Common Income Supports Service



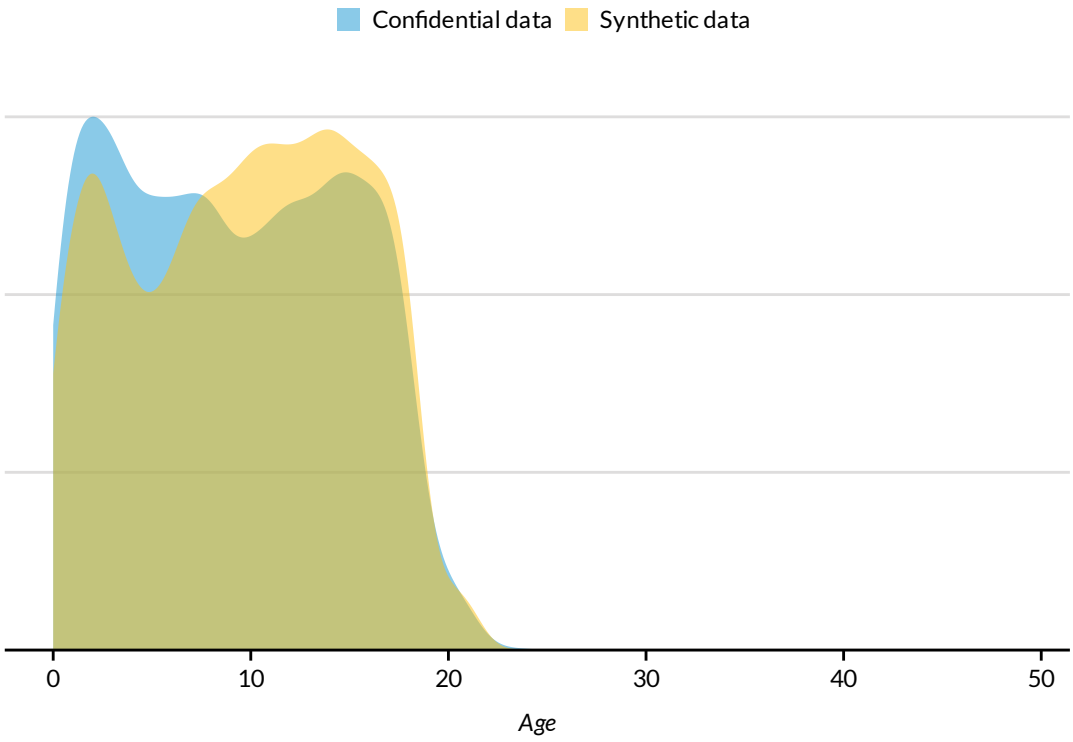
URBAN INSTITUTE

Source: 2021 DHS confidential data and 2021 Integrated Services data

Figure 5B compares the confidential and synthetic age distributions for recipients of the child welfare service. This service is much less common and is synthesized later in the synthesis order, so there is more overall error propagated through to these synthesized values. While the distributions do not closely align at each point, the synthetic data still mimicked the general shape of the confidential distribution.

FIGURE 5B

Synthetic Data Captures the General Shape of the Age Distribution for the Child Welfare Service



URBAN INSTITUTE

Source: 2021 DHS confidential data and 2021 Integrated Services data.

Discriminant-based metrics. Finally, we investigated how well a classification model could differentiate between records in the confidential and synthetic data; due to the categorical nature of the data, applying a tree-based model in this context was a natural extension of our synthesis methodology. For greatest utility, we want this model to perform poorly—indicating that the synthetic and confidential data have such similar statistical properties that the model has difficulties discriminating between them.

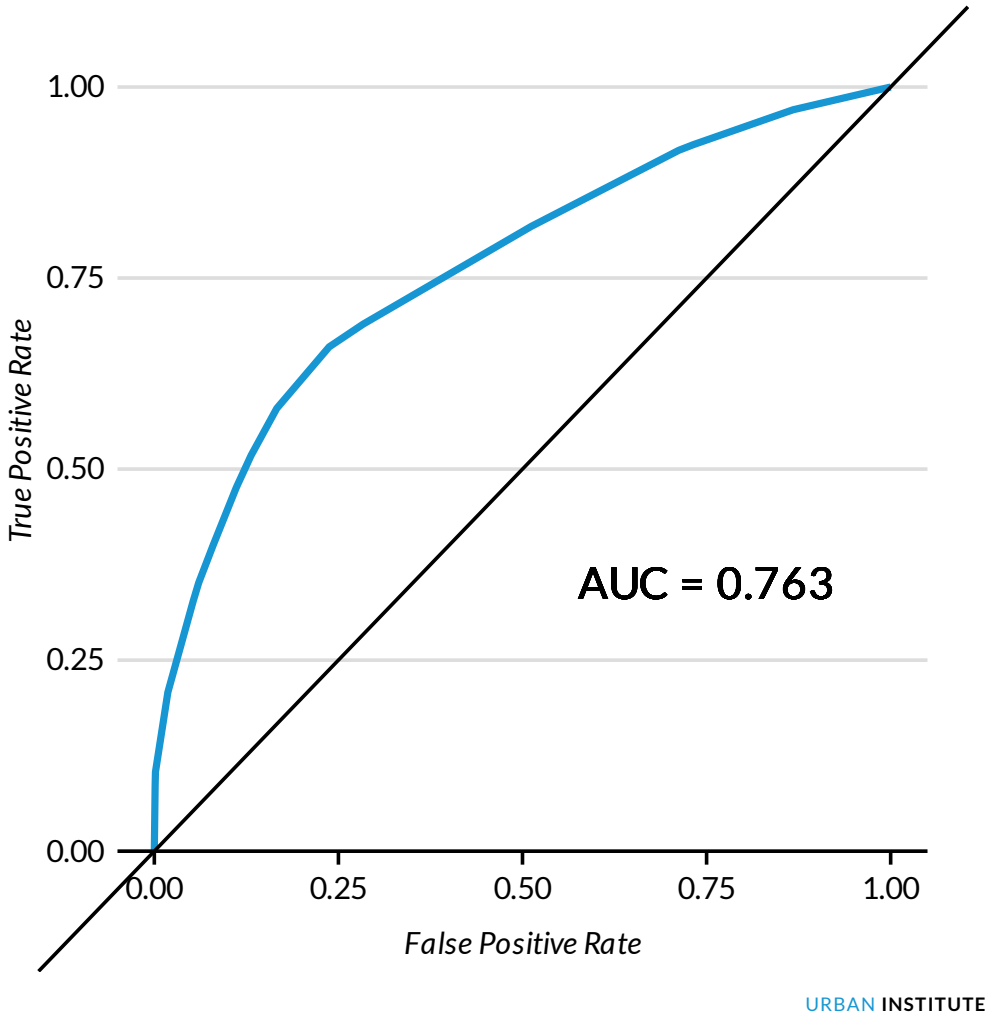
We used a classifier model to generate a propensity score for each record (i.e., the probability that the record belongs to the confidential data) and used these propensity scores to plot a receiver operating characteristic (ROC) curve. This curve enabled us to visualize the model's discriminating performance as we varied the classification threshold. According to this metric, the synthetic data most similar to the confidential data would have an ROC curve close to the 45-degree line and the resulting area under the curve (AUC) close to 0.5. This would indicate that the discriminator cannot distinguish between the confidential and synthetic data (Mendelevitch and Lesh 2021).

We trained the discriminant model on the entire combined dataset and then evaluated the performance of the model on the same dataset using the resulting ROC curves and AUC values. We also evaluated the sensitivity of results to the size of the classification trees by varying the complexity parameter value. Figure 6 shows the ROC and AUC for the selected synthetic dataset using a complexity parameter value of 0.005 in the classification model.

The selected synthetic dataset had a relatively low AUC value, compared with other candidate synthetic data sets, indicating that these synthetic records are generally more difficult to distinguish from the confidential records. We explored variable importance to investigate which variables were most important for the model to be able to distinguish between the confidential and synthetic data and found two primary takeaways. First, age, education level, and marital status were the most important variables. This is reflective of the postprocessing steps we applied, discussed earlier in this report. While we determined with our county partners that these postprocessing steps were important to preserve realistic values in the synthetic dataset, these steps did somewhat reduce data utility as measured by these discriminant metrics (i.e., a slightly higher AUC value). Second, we found high variable importance for the income supports service in months July through December, which is reflective of the additional error propagated to later months of the year as a result of our two-step synthesis process.

FIGURE 6

A Classification Model Shows Moderate Ability to Differentiate between Confidential Records and Synthetic Records from the Selected Synthetic Dataset



Source: 2021 DHS confidential data and 2021 Integrated Services data

Evaluating Synthetic Data Privacy

Another key dimension for evaluating the synthetic dataset is disclosure protection. While individual-level or record-level data are useful for powerful and flexible analyses, the release of such disaggregated data could expose participants to disclosure risks. Increases in computing power and public releases of supplementary data could empower malicious actors to seek to re-identify the individuals whose data were synthesized and disclose sensitive information about them. We worked

closely with county partners to evaluate these risks in the specific context of the synthetic 2021 Integrated Services dataset and to determine alignment with county privacy policies.

Disclosure Risk Considerations

We view identity and attribute disclosure risks associated with the release of the 2021 Synthetic Integrated Services dataset as low, though inferential disclosure risk remains a possibility. Because the dataset is fully synthetic, there is no one-to-one mapping between synthetic and confidential records. Consequently, data users have no way of accurately associating records in the synthetic datasets with real individuals. This makes both identity and attribute disclosure risk highly unlikely (Raab, Nowok, and Dibben 2017; Hu, Reiter, and Wang 2014; Reiter 2002; Bowen et al. 2020).

In the Data Synthesis Methodology section, we described our exclusion of sparse and potentially more sensitive variables from the GSDS (e.g., geographic area, sexual orientation, and veteran status). These exclusions serve as an additional layer of disclosure control because the inclusion of these variables could have provided additional identifying information to a malicious actor. Further, reducing the number of sparse groups that would allow for precise but uncertain inferences protects specifically against inferential disclosure risk.

It is important to note that attribute disclosure risk could be considered nontrivial if a malicious actor has meaningful information about the records in the confidential data or the modeling process used to generate the synthetic records. However, in this case, such risks seem unlikely because the county has not disseminated the confidential data publicly, and we only share limited details about our synthetic data generating process. Moreover, even with significant information, an attacker's confidence in an attempted attribute attack would be limited because they cannot confirm which values exist in the confidential data (Bowen et al. 2020).

We therefore focus on assessing the risk that our modeling process could have generated synthetic data that aligns too closely with the confidential data. This could happen if the data generating model is overfit to the confidential data, resulting in significant copying of confidential records in the synthetic data (Bowen et al. 2020; Elliot 2014). Further, this type of copying risk is particularly salient given the largely categorical nature of the confidential dataset. There are a finite number of combinations of categorical variables that have just a few values (Bowen et al. 2020), and certain combinations of variable values will have greater likelihood of occurring if there are class imbalances. The metrics we use to evaluate these risks are described in the following section. To maintain data privacy, we do not

share the results of these assessments publicly, but we did share and discuss them with our county partners.

Disclosure Metrics

We explored three sets of metrics to evaluate disclosure risks: duplicates, unique-uniques, and distance to closest record. These metrics specifically enabled us to assess risks related to dataset copying. We discussed these findings with the county partners to ultimately select a synthetic candidate dataset with sufficient privacy protections, but do not report results to minimize privacy risks.

Duplicates. We first evaluated the extent that records were duplicated across the confidential and synthetic datasets. Although no actual records appeared in the synthetic data, Bowen and colleagues (2020) note that "a row in the synthetic data could match a row in the confidential data by chance." We examined the counts and shares of total and distinct records in the synthetic data that exactly matched records in the confidential data.

The confidential data are largely categorical, and the categorical variable classes are sometimes very imbalanced. For example, while there are seven possible race categories, 93 percent of records are either White or Black/African American. Further, some service receipt patterns were much more common than others. For example, the majority of records received only the income supports service in each month and no other services. Further, the synthetic dataset only includes only six demographic variables with a finite number of combinations. We expected record duplication in this context, and this type of duplication does not necessarily increase disclosure risk (Bowen et al. 2020).

In consultation with the county partners, we determined that record duplication could be more concerning for disclosure risk if certain populations were subjected to higher duplication rates than other populations. To evaluate this, we compared the demographic characteristics of duplicated records with the demographic characteristics of the total universe of records and found no concerning differences.

Unique-uniques. We further investigated duplication and copying risks by examining "unique-uniques," or unique records in the confidential data that are also unique in the synthetic data. These are uncommon records that could carry higher disclosure risks (Bowen et al. 2020).

Unique-uniques are less common than the more general duplicated values but still can be expected to appear by chance as a result of the synthesis process and the nature of the underlying data. As with duplicates, we compare the characteristics of unique-uniques with those of unique and total

confidential records to determine if certain populations may be subjected to higher disclosure risks. We found no concerning differences.

Distance to closest record. The distance to closest record (DCR) is the distance between a confidential record and the synthetic record that is closest to it (Mendelevitch and Lesh 2021). Duplicated records have DCR = 0, while records with only simple perturbations have a small DCR value and those with large perturbations have a large DCR value.

Generally speaking, a synthetic dataset with a high level of privacy preservation would have very few zero or small DCR values, indicating that few confidential records appear in the synthetic dataset as duplicates or with only small perturbations. However, Mendelevitch and Lesh (2021) note that DCR = 0 does not necessarily indicate high disclosure risks when the "'space' spanned by the variables in scope is relatively small," which can occur with largely categorical data with class imbalances.

We used Gower distance, a common distance metric for data with numeric and categorical variables (Gower 1971), to find the DCR for each confidential record and then evaluate the distribution of these distances.

We found that DCR results were generally not meaningful in this context and face significant computational constraints when running these analyses, so we focused our evaluation primarily on "high-risk records," or those with DCR = 0 (duplicates) that have five or fewer copies within the confidential data and compared the share of these records with common benchmarks (Mendelevitch and Lesh 2021). Again, we also compared the characteristics of "high-risk records" with those of all duplicate and all confidential records to determine if certain populations were subjected to higher disclosure risks and found no concerning differences.

Conclusions and Future Work

Our final synthetic product was heavily informed by the use case we defined with our partners at the beginning of this work: allowing data users to work with disaggregated, longitudinal service data without needing to apply for confidential data access. Centering this use case informed each stage of our dataset creation: the structure of the GSDS as a modeling input, the two stages of modeling we used to synthesize the dataset, and the postprocessing and privacy-quality evaluation that we conducted. Defining this use case was thus arguably the most crucial step in our work.

Another key theme throughout the creation of this dataset was the iterative nature of the work. Evaluating each candidate synthesis for privacy and quality led to new conversations about which aspects of the data were most important to preserve, how we could best structure the data to facilitate that, how we could measure quality and privacy risk of the synthetic output, and what acceptable levels of error and privacy risk should be for the final selection. We depended heavily on input from our partners as domain experts when making these decisions.

The computational intensity and bespoke nature of the synthetic data generation process could pose obstacles to many local governments. For this project, Urban Institute data scientists spent more than 250 total hours working to develop and assess the synthetic data. They also used cloud computing resources because of the computational intensity of their work. Many local governments may not have staff with the technical capabilities to perform similar work or the budget to fund expensive cloud computing operations. While Allegheny County DHS and WPRDC staff are highly technical, Urban's involvement on the project was key to provide staff capacity and data privacy domain knowledge to complete this work. Relatedly, early in the project, Urban considered trying to develop a tool to automatically generate synthetic data from an uploaded dataset. However, given the highly specific nature of synthetic data to a specific use case, we quickly abandoned that idea as infeasible. The fact that synthetic data must be bespoke further adds to the capacity constraints facing local governments who may seek to do similar work. While the process could be sped up for datasets with similar structure, for example by repeating the synthesis process on a different year of the same data, it would always require some level of manual review and input before approval. Education and training for local government staff will be key to minimizing these constraints and making synthetic data more accessible in the local context.

Synthetic data presents a unique opportunity for local governments and organizations to release disaggregated administrative data while protecting individual privacy. Users of this dataset will be able

to answer questions that were not previously accessible with the publicly available data. We hope that this work will lay the foundation for future disaggregated synthetic datasets and encourage data curators to consider additional privacy-protecting methods when releasing data.

Notes

[1] We use the term "synthetic individual" to refer to an entirely theoretical person represented in our synthetic dataset by all records with the same Client Identifier. The term "synthetic individual" is not common in the data synthesis literature. Many scholars use "synthetic records," to refer to rows in their synthetic datasets. However, in our synthesis process, we generated a synthetic dataset that has multiple records (rows) that have the same client identifier. We have done this intentionally to mimic the structure of the confidential dataset. Because we want to refer to the theoretical individual not just a record reflecting a service they received in a specific month, we use the term "synthetic individual."

References

Allegheny County Department of Human Services. 2021. "Allegheny County Human Services Community Profiles." Pittsburgh, PA: Western Pennsylvania Regional Data Center.

Bowen, Claire McKay. 2021. "Personal Privacy and the Public Good." Washington, DC: Urban Institute.

Bowen, Claire McKay, Aaron Williams, and Madeline Pickens. 2022. "Decennial Disclosure." Washington, DC: Urban Institute.

Bowen, Claire McKay, Victoria L. Bryant, Len Burman, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Philip Stallworth, Kyle Ueyama, Aaron R. Williams, and Noah Zweifel. 2020. "A Synthetic Supplemental Public-Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications." Washington, DC: Urban Institute.

Drechsler, Jörg, and Jingchen Hu. 2021. "Synthesizing Geocodes to Facilitate Access to Detailed Geographic Information in Large-Scale Administrative Data." *Journal of Survey Statistics and Methodology* 9.3 (2021): 523–48. https://doi.org/10.1093/jssam/smaa035.

Elliot, Mark. 2014. "Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team." Manchester, UK: University of Manchester.

Fellegi, Ivan P. 1972. "On the Question of Statistical Confidentiality." *Journal of the American Statistical Association* 67 (337): 7–18. https://doi.org/10.1080/01621459.1972.10481199.

Gower, John C. 1971. "A general coefficient of similarity and some of its properties." *Biometrics*: 857-871. https://doi.org/10.2307/2528823.

Hu, Jingchen, Jerome P. Reiter, and Quanli Wang. 2014. "Disclosure Risk Evaluation for Fully Synthetic Categorical Data." In: Domingo-Ferrer, J. (eds). Privacy in Statistical Databases. PSD 2014. Lecture Notes in Computer Science, vol 8744. Springer, Cham. https://doi.org/10.1007/978-3-319-11257-2_15.

Mendelevitch, Ofer and Michael D. Lesh. 2021. "Fidelity and Privacy of Synthetic Medical Data: Review of Methods and Experimental Results." San Francisco, CA: Syntegra.

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2017. "Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R." Statistical Journal of the IAOS 33: 785–796. https://doi.org/10.3233/sji-150153.

Reiter, Jerome P. 2002. "Satisfying Disclosure Restrictions with Synthetic Datasets." Journal of Official Statistics 18: 531–543.

Therneau, Terry, and Elizabeth Atkinson. 2022. "An Introduction to Recursive Partitioning Using the RPART Routines." Available at https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Williams, Aaron R. 2022. "The tidysynthesis R package." Presentation given at rstudio::conf(2022), Washington, DC, July 25 – 28.

Woolley, Michael E., Laura M Stapleton, Daniel Bonnéry, Mark Lachowicz, Terry V. Shaw, Angela K. Henneberger, Bess A. Rose, Tessa L. Johnson, Yi Feng. 2020. "Expanding MLDS Data Access and Research Capacity with Synthetic Datasets." Baltimore, MD: Maryland Longitudinal Data System Center.

About the Authors

Madeline Pickens is a data scientist at the Urban Institute. She works with Urban's Technology and Data Science team to support research and analysis relating to data privacy.

Before joining Urban, she worked as a data analyst at the Universal Service Administrative Company on improving broadband access. She also worked as a graduate fellow in the University of Virginia's Social and Decision Analytics Division.

Madeline holds a bachelor's degree in economics from the University of Arizona, where she was a Flinn Scholar, and a master's degree in data science for public policy from Georgetown University, where she was a Whittington Scholar.

Jennifer Andre is a data scientist in the Center on Labor, Human Services, and Population at the Urban Institute, focusing on financial well-being research.

Before joining Urban, she worked as an economic consulting analyst at Charles River Associates in the antitrust and competition practice.

Andre holds a BA in economics from the University of Notre Dame and an MS in public policy and management–data analytics from Carnegie Mellon University.

Gabriel Morrison is a data scientist on the Technology and Data Science team at the Urban Institute. His work focuses on researching and building tools to address equity-related issues in cities and counties, often with spatial data. Morrison also helps manage and govern Urban's use of cloud computing resources.

Before joining Urban full time, Morrison interned with the Cook County Assessor's Office, Urban's Data Science team, the University of Chicago's Center for Spatial Data Science, the Brookings Institution's Metropolitan Policy Program, the Metropolitan Planning Council, and Taller de José.

Morrison holds a BA in geographical sciences and an MS in computational analysis and public policy from the University of Chicago.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org