# Synthetic Data User Guide

Madeline Pickens, Jennifer Andre, and Gabe Morrison

*August 2023*

## Introduction

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. These data are integrated at the individual level to allow for better care coordination, operational improvements, and program evaluation. Because of the dataset's sensitive nature, it cannot be widely shared at an individual level, so synthetic data are used in the real dataset's place—allowing the data to be publicly shared and helping stakeholders, including researchers, service providers, and members of the public, understand these populations better.

This user guide accompanies a fully synthetic version of the 2021 Integrated Services dataset, which replaces the underlying records tracking the use of these services with statistically representative pseudo-records. Each record in the synthetic dataset represents a simulated individual, or record, who received at least one service from the Allegheny County DHS in 2021. The synthetic data were designed such that records aggregated by service represent the original data.

To create this synthetic data product, staff at the Urban Institute partnered with the Allegheny County DHS and the Western Pennsylvania Regional Data Center (WPRDC). The Urban Institute has a body of work dedicated to data privacy and has previously created synthetic datasets at the federal level. This partnership is intended to function as a pilot for synthetic data generation at the local level, to help understand the unique challenges that might face state and local governments in generating synthetic data. As this is the first synthetic data product released by Allegheny County and the WPRDC, this guide is intended to provide an overview of the motivation behind the creation of a synthetic version of this dataset, a high-level summary of the data synthesis process, and information that will allow users to make informed decisions while using this dataset.

## Why Synthetic Data?

The Allegheny County DHS collects administrative data about service usage for the purpose of care coordination, case management, and quality improvement efforts. While these data are released publicly, they are aggregated, or grouped on specific features, to protect individual privacy. The level of aggregation is such that any geographies with fewer than six individuals represented are suppressed (i.e., the values are not reported publicly).

The publicly released aggregated data contrast with the Allegheny County DHS's private disaggregated data, which comprise individual-level records with information about a particular service recipient and the services they received. Each record reflects that an individual received a service at least once a month, so the data are disaggregated.

Although aggregating and suppressing data can protect individual privacy, these measures limit potential applications of the service use data. Interactions between the usage of different services and the repeated nature of the service distribution are not captured. For example, the existing public data would not allow a data user 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.

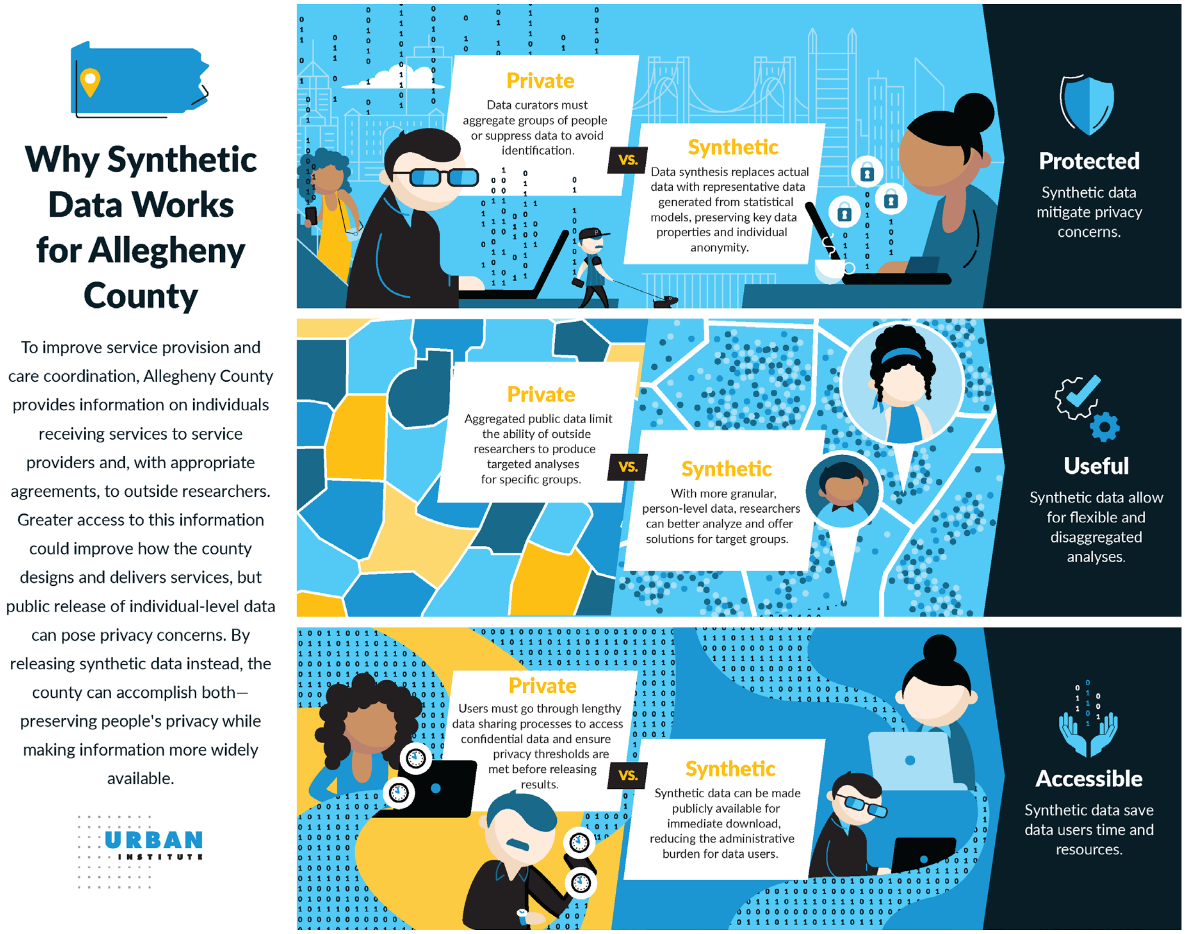
Additionally, if the data are aggregated or suppressed, some types of analyses are impossible. For example, consider a researcher studying the effect of service provision on upward mobility. This researcher could investigate whether an individual receiving income support and other services at the beginning 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. However, disaggregated synthetic data could allow for this analysis.

Generating synthetic data is a statistical technique that allows for the release of disaggregated data while mitigating some of the threats to individual privacy. A team of Urban Institute researchers generated each record in the synthetic Integrated Services dataset from models that are representative of the confidential data. With record-level synthetic data, users can begin to understand the interactions among the usage of different services.

Not all elements of the confidential dataset can be preserved through the synthesis process to ensure privacy protections remain robust. If data users find that their desired analysis is still not possible with the synthetic data, they may request the confidential data from the county through existing processes. Even if data users hope to eventually gain access to the county's private data, the released synthetic data can provide initial guidance regarding the structure of the dataset, such as the variable types and value ranges. This initial data access will make the process more efficient for stakeholders to identify the specific datasets within the county's database they want to access. Users can also rely on the synthetic dataset as training data to run and debug their analyses faster once access to the confidential data is granted.

FIGURE 1

**Why Synthetic Data Works for Allegheny County**



Why Synthetic Data Works for Allegheny County

To improve service provision and care coordination, Allegheny County provides information on individuals receiving services to service providers and, with appropriate agreements, to outside researchers. Greater access to this information could improve how the county designs and delivers services, but public release of individual-level data can pose privacy concerns. By releasing synthetic data instead, the county can accomplish both—preserving people's privacy while making information more widely available.

Private: Data curators must aggregate groups of people or suppress data to avoid identification.

VS.

Synthetic: Data synthesis replaces actual data with representative data generated from statistical models, preserving key data properties and individual anonymity.

Protected: Synthetic data mitigate privacy concerns.

Private: Aggregated public data limit the ability of outside researchers to produce targeted analyses for specific groups.

VS.

Synthetic: With more granular, person-level data, researchers can better analyze and offer solutions for target groups.

Useful: Synthetic data allow for flexible and disaggregated analyses.

Private: Users must go through lengthy data sharing processes to access confidential data and ensure privacy thresholds are met before releasing results.

VS.

Synthetic: Synthetic data can be made publicly available for immediate download, reducing the administrative burden for data users.

Accessible: Synthetic data save data users time and resources.

URBAN INSTITUTE

## About the Data

The synthetic dataset that accompanies this guide tracks individual-level usage of all DHS services throughout 2021. The dataset is structured on the service-month level, meaning that each row in the dataset is associated with one synthetic individual, one service received by that individual, and the month in 2021 that the individual received the service. If a given synthetic individual received multiple services, one service in multiple months, or some combination, they would be represented by more than one row in the dataset.

Although many of the services are provided more frequently than once a month, Allegheny County tracks service usage monthly as a baseline for comparison across the services. The synthetic dataset treats the month variable in the same way.

The service variable could take one of 22 possible values, each representing the particular service that was provided in a given month. These possible values are 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[*]

As many individuals receive services in multiple months and multiple services in 2021, multiple rows in the dataset may be associated with one service recipient. An ID variable serves as a unique identifier of service recipients. Note that this ID does not have any association with the ID from the confidential data, but it can tie repeat service usage to one synthetic individual. Demographic variables included in each row are race, ethnicity, gender, marital status, educational status, and living arrangement. These demographic values remain the same across every record associated with one synthetic individual. Demographic data represent the most recent information for an individual in June 2022, when the data were pulled from the DHS system.

For exact variable names and additional context for the service variable, including criteria for receiving a given service, see the accompanying data dictionary.

### *Excluded Variables*

Some demographic fields in the confidential data are missing values for most service recipients. For example, gender identity is missing for the vast majority of DHS service recipients. Additionally, in some cases, these sparsely populated fields can be correlated with usage of particular services. Therefore, even synthesized values for these variables could pose a disclosure risk. Further, these variables would

[*] Although "services" is used throughout this guide for brevity, these variables denote *incidents* to which Allegheny County DHS responded.

be difficult to replicate in a synthesis because of their sparsity, so any additional information they provide to data users would be limited. For these reasons, these variable values are excluded from the synthetic output.

The geographic area variable, which lists the municipality where the service recipient lived, is also excluded. Despite available data, this variable is excluded from the synthetic product because many municipalities had very few associated individuals. The records would have to be suppressed to align with county policy, and the variable distribution would be difficult to replicate in the synthesis without sacrificing other elements of data quality deemed more important for this particular product, namely the relationship between the usage of different services and service usage over time. Users interested in the geographic elements of this data are encouraged to review the published aggregate statistics or request access to the underlying data.

Finally, the date of birth and date of death variables are omitted because of privacy concerns, though a synthetic age variable reflects the age of the service recipient in 2021.

Although meaningful values are excluded from these variables, the synthetic product includes columns for each of these variables, populated with “N/A” (null), to give users an understanding of the structure of the confidential data. Ultimately, this applies to the following variables:

- Geographic area
- Date of birth (replaced with age)
- Date of death
- Gender identity
- Sexual orientation
- Legal sex
- Employment status
- Living situation
- Veteran status

## Creating the Synthetic Dataset

All data synthesis requires some level of iteration to identify the optimal balance between data utility and data privacy. Data curators (individuals or entities responsible for safeguarding an organization’s data) must balance the privacy risk posed by data releases with the assurance the synthetic product is as useful and representative of the confidential data as can be achieved. Many separate synthetic datasets were generated using variations on analytical decisions, such as the models used, input data, and order in which variables were synthesized. We evaluated these dataset options for data quality and risk of privacy violations before ultimately selecting the version for publication.

For more information about the creation of this dataset, see the technical report for this project, which discusses the technical decision making and modeling process in more detail.

### *Generating Synthetic Values*

Allegheny County DHS provided separate individual-level data for each service, so the synthesis process began by combining these separate sources into a single confidential dataset linking individual-level data across 22 services provided by Allegheny County in 2021. We trained a series of models on this dataset to capture the relationships among the variables, and we used these models to generate new variables sequentially, with previously synthesized variables used as predictors. Because completely new values were generated for all records, the new synthetic dataset constitutes fully synthetic data.

### *Applying Constraints*

We applied constraints during the synthesis process to prevent unrealistic age values given the human services context. For example, DHS-funded out-of-school programs are only available to children under 18, so the models did not allow individuals over 18 to be assigned this service.

### *Post-Processing*

Some records for children in the synthetic output contained unrealistic values for demographic variables, namely marital status and education status. We addressed these values by converting any unrealistic values to "unknown." Some examples include children under 16 with an education level higher than high school and children under 18 with a marital status other than "unknown" or "single, never married."

## Evaluating the Synthetic Dataset

After generating many candidate synthetic datasets through the process described above, we selected the candidate synthesis by evaluating its quality and privacy risk in alignment with the proposed use case and county priorities and policies. This section describes the metrics used to assess quality and privacy risk, explains the logic behind selecting those metrics, and outlines limitations of the dataset. For more detail on any of the metrics or processes described below, see the technical report for this project.

### *Evaluating Quality*

We applied a variety of metrics and comparisons to evaluate the quality of the synthesis versions. The quality of a synthetic dataset is highly dependent on the dataset use case. Allegheny County publicly releases aggregated human service data at the municipality scale. Consequently, the preservation of key demographic characteristics, patterns around receipt of multiple services, and service receipt across 2021 are the most important aspects of the confidential dataset to maintain. Metrics for evaluating these goals are discussed at a high level below.

Note that it is impossible to perfectly replicate all features of the confidential dataset; however, checking that the features are broadly replicated ensures the utility of the synthetic product.

We used the following general utility metrics to evaluate synthetic data quality:

- **Comparisons of categorical variable value frequencies**. For each categorical variable separately and in various combinations, we compared the frequency of certain categories or combinations of categories in the synthetic data with the confidential data to determine how well the synthetic data capture these values.
- **Comparisons of person-level and monthly service counts.** We compared the distributions of the number of services received by individuals in the confidential and synthetic data as well as the total counts of service receipt by month in total and separately by service. These metrics indicate how well the synthetic data capture the usage of different services and the usage of services over time.
- **Comparisons of numeric variable summary statistics and distributions.** For the age variable, we compared the summary statistics (e.g., mean, median, variance) and distributions overall and by service between the confidential and synthetic data to determine how well the synthetic data capture relationships between age and service use.
- **Discriminant-based metrics.** Machine learning algorithms assessed the extent to which confidential and synthetic records were distinguishable. Poor performance of an algorithm would indicate that the records in the confidential and synthetic data are more similar to each other, and, thus, the synthetic data were of higher quality. The selected synthetic dataset had high performance across many of these metrics relative to other candidate options.

### *Evaluating Privacy Risk*

Although our synthetic file should reflect the confidential file as well as possible, a tradeoff exists between the quality of the synthesis and the privacy protections provided, so evaluating the disclosure or privacy risk associated with the chosen dataset is also necessary. There are three main types of disclosure risks, or threats to privacy resulting from the release of a dataset or statistic: The first type, identity disclosure, refers to the association of a specific individual with a released record. The second type, attribute disclosure, occurs if an attacker can determine characteristics of an individual based on the information in the released data. The final type, inferential disclosure, occurs if an attacker can predict the value of some individual characteristic more accurately with the public data or statistic than would otherwise have been possible.

The synthetic services dataset is fully synthetic, meaning no one-to-one mapping between the confidential and synthetic data exists. For fully synthetic data, identity and attribute disclosure risks are considered low. Even with significant information about the underlying data and synthesis process, an attacker's attempt to disclose sensitive information is limited without direct links to the confidential data. Our exclusion of sparse variables protects against inferential disclosure by reducing the number of small groups that would allow for precise but uncertain inferences.

Given the low risks of identity, attribute, and inferential disclosure in this context, the primary focus shifts to assessing the risk that our modeling process could have generated synthetic data that aligns too closely with the confidential data. This can occur when the synthesis process essentially results in direct copies of the confidential data. We used three primary metrics to explore this type of disclosure risk:

- **Duplicates.** How many records are duplicated across both the confidential and synthetic datasets?
- **Unique-uniques.** How many records that appear only once (i.e., they are unique) in the confidential data are also unique in the synthetic data?
- **Distance to Closest Record.** For each record in the confidential data, how similar is the most similar record in the synthetic data?

For each of these metrics, we evaluated demographic distributions of affected records to ensure that no specific group had a meaningfully higher probability of being perfectly replicated and having potentially higher disclosure risk. Like all data products publicly released by the WPRDC, the synthetic data underwent a final disclosure review to ensure its compliance with county policy from a privacy standpoint.

## Limitations

Although providing exact benchmarks using the confidential data can undermine the privacy protections of synthetic data, some of the known limitations of the synthetic data are discussed below. Note that additional limitations to this dataset not captured in the evaluation process exist.

The overall distribution of service counts received by each person matches well. As expected, there is clustering around multiples of 12, because many individuals receive a service (or more than one) at least once a month every month of the year. The synthetic data slightly undercount individuals who received 24 and 36 service-months and slightly overcount individuals who received more than 12 service-months, but not a multiple of 12. In general, earlier months in the year (January through March) tended to overcount service receipt, while service receipt in December was undercounted.

Some populations are overrepresented in the synthetic data. These populations include the jailed population, families receiving Child Welfare services, the homeless population, those who have experienced homicide and suicide, and those who have overdosed. Older adults receiving services are undercounted in the synthetic data. In all cases, despite an overcount or undercount in a given month, the shape of the distribution remains similar across the synthetic and confidential data. For example, a drop in recipients from June to July would be reflected in the synthetic data, even if the synthetic data are overrepresenting the number of individuals in both June and July.

The synthetic dataset tends to slightly undercount younger individuals receiving services. Exceptions to this are synthetic individuals in the homeless population and those who have experienced suicide. These services are overcounted for individuals under 18 in the synthetic data. Service usage by demographic was also compared across the synthetic and confidential data. The tables below list service-demographic combinations that exceeded a 20 percent difference from the confidential data for race, ethnicity, and gender. These differences affected a relatively small number of service-demographic combinations, and primarily affected services with a small sample size, making the differences more pronounced when represented as a percentage.

TABLE 1

### Under- and Overrepresentation of Gender in the Synthetic Data

| Service | Representation in synthetic data |
|---|---|
| *Overdoses* | |
| Male | Underrepresented by at least 20% |
| Female | Overrepresented by at least 20% |
| *Suicides* | |
| Male | Underrepresented by at least 20% |
| Female | Overrepresented by at least 20% |

**Source:** Comparison of confidential and synthetic datasets.

**Note:** The exact level to which demographics are under- or overrepresented is not reported for privacy reasons.

TABLE 2

### Under- and Overrepresentation of Race in the Synthetic Data

| Service | Representation in synthetic data |
|---|---|
| *Children Attending Early Childhood Programs Managed by DHS* | |
| White | Overrepresented by at least 20% |
| Black/African American | Underrepresented by at least 20% |
| *Children Receiving Early Intervention Services* | |
| Unknown | Underrepresented by at least 20% |
| *Individuals Receiving Family Strengthening Programs* | |
| White | Overrepresented by at least 20% |
| Black/African American | Underrepresented by at least 20% |

**Source:** Comparison of confidential and synthetic datasets.

**Note:** The exact level to which demographics are under- or overrepresented is not reported for privacy reasons.

TABLE 3

### Under- and Overrepresentation of Ethnicity in the Synthetic Data

| Service | Representation in synthetic data |
|---|---|
| *Children Attending Early Childhood Programs Managed by DHS* | |
| Not Hispanic/Latinx | Underrepresented by at least 60% |
| *Children Receiving DHS Funded Out of School Programs* | |
| Unknown | Overrepresented by at least 40% |
| *Children Receiving Early Intervention Services* | |
| Not Hispanic/Latinx | Overrepresented by at least 20% |
| Unknown | Underrepresented by at least 30% |
| *Homicides* | |
| Not Hispanic/Latinx | Underrepresented by at least 20% |
| Unknown | Overrepresented by at least 20% |
| *Overdoses* | |
| Not Hispanic/Latinx | Underrepresented by at least 20% |
| Unknown | Overrepresented by at least 20% |
| *Suicides* | |
| Not Hispanic/Latinx | Underrepresented by at least 40% |
| Unknown | Overrepresented by at least 50% |

**Source:** Comparison of confidential and synthetic datasets.

**Note:** The exact level to which demographics are under- or overrepresented is not reported for privacy reasons.

***Madeline Pickens** is a data scientist in the Office of Technology and Data Science at the Urban Institute. Her research focuses on applications of data science methodology in data privacy.*

***Jennifer Andre** is a data scientist in the Center on Labor, Human Services, and Population. Her research focuses primarily on financial well-being.*

***Gabe Morrison** is a data scientist in the Office of Technology and Data Science. His research focuses on visualization and analysis of spatial data.*

## Acknowledgments

This user guide 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 user guide 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 author 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 www.urban.org/fundingprinciples.

URBAN INSTITUTE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org

### ABOUT THE URBAN INSTITUTE